DOCUMENT RESUME

ED 164 303 SE 025 458

AUTHOR Beck, A.; And Others
TITLE Calculus, Part 3, Student's Text, Unit No. 70. Revised Edition.
INSTITUTION Stanford Univ., Calif. School Mathematics Study Group.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE 65
NOTE 360p.; For related documents, see SE 025 456-459; Contains light and broken type

EDRS PRICE MF-$0.83 HC-$19.41 Plus Postage.
DESCRIPTORS *Calculus; *Curriculum; *Instructional Materials; *Mathematical Applications; Mathematics Education; Secondary Education; *Secondary School Mathematics; *Textbooks
IDENTIFIERS *School Mathematics Study Group

ABSTRACT

This is part three of a three-part SMSG calculus text for high school students. One of the goals of the text is to present calculus as a mathematical discipline as well as presenting its practical uses. The authors emphasize the importance of being able to interpret the concepts and theory in terms of models to which they apply. The text demonstrates the origins of the ideas of the calculus in practical problems; attempts to express these ideas precisely and develop them logically; and finally, returns to the problems and applies the theorems resulting from that development. Chapter topics include: (1) vectors and curves; (2) mechanics; (3) numerical analysis; (4) sequences and series; and (5) geometricl optics and waves. (MP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH.
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SMSG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED164303

SE 025 458

Calculus

Part 3 Student's Text

REVISED EDITION

The following is a list of all those who participated in the preparation of this volume:

A. Beck	Olney High School, Philadelphia, Pa.
A. A. Blank	New York University, New York, N.Y.
F. L. Elder	West Hempstead Jr., Sr. High School, N.Y.
C. E. Kerr	Dickinson College, Carlisle, Pa.
M. S. Klamkin	Ford Scientific Laboratory, Dearborn, Mich.
I. I. Kolodner	Carnegie Institute of Technology, Pittsburgh, Pa.
M. D. Kruskal	Princeton University, Princeton, N.J.
C. W. Leeds, III	Berkshire School, Sheffield, Mass.
M. A. Linton, Jr.	William Penn Charter School, Philadelphia, Pa.
H. M. Marston	Douglass College, New Brunswick, N.J.
I. Marx	Purdue University, Lafayette, Ind.
R. Pollack	New York University, New York, N.Y.
T. L. Reynolds	College of William and Mary, Williamsburg, Va.
R. L. Starkey	Cubberley High School, Palo Alto, Calif.
V. Twersky	Sylvania Electronics Defense Labs., Mt. View, Calif.
H. Weitzner	New York University, New York, N.Y.

Stanford, California
Distributed for the School Mathematics Study Group
by A. C. Vroman, Inc., 367 Pasadena Avenue, Pasadena, California

Financial support for School Mathematics
Study Group has been provided by the
National Science Foundation.

Permission to make verbatim use of material in this book must be secured from the Director of SMSG. Such permission will be granted except in unusual circumstances. Publications incorporating SMSG materials must include both an acknowledgment of the SMSG copyright (Yale University or Stanford University, as the case may be) and a disclaimer of SMSG endorsement. Exclusive license will not be granted save in exceptional circumstances, and then only by specific action of the Advisory Board of SMSG.

© 1965 by The Board of Trustees
of the Leland Stanford Junior University.
All rights reserved.
Printed in the United States of America.

TABLE OF CONTENTS

Chapter 11. VECTORS AND CURVES 669

11-1. Introduction 669
11-2. Vector Algebra 672
11-3. Vector Geometry 681
11-4. Products of Two Vectors. 690
11-5. Vector Calculus and Curves 704
11-6. Curves in the Plane. 719
Miscellaneous Exercises 743

Chapter 12. MECHANICS. 747

12-1. Introduction 747
12-2. Elementary Mechanical Problems 755
12-3. Constraints. Use of Energy Conservation 772
12-4. Angular Momentum and Central Forces. 790
Miscellaneous Exercises. 803

Chapter 13. NUMERICAL ANALYSIS 805

13-1. Introduction 805
13-2. Iteration 808
13-3. Taylor's* Theorem with Remainder 819
13-4. Numerical Integration 829
13-5. Numerical Solution of First Order Differential Equations 842
Miscellaneous Exercises 848

Chapter 14. SEQUENCES AND SERIES 851

14-1. Introduction 851
14-2. Convergence of Sequences 852
14-3. Series 863
14-4. Conditional and Absolute Convergence 873
14-5. Parentheses and Rearrangements 877
14-6. Sequences of Functions, Uniform Convergence 883
14-7. Power Series 891
Miscellaneous Exercises 896

Chapter 15. GEOMETRICAL OPTICS AND WAVES 901

15-1. Introduction 901
15-2. Geometrical Optics 903
15-3. Refraction 927
15-4. Kepler-Lambert Principle 939
15-5. Huyghen's Principle. 950
15-6. Periodic Waves 959
15-7. Method of Stationary Phase 977
15-8. Mathematical Model for Scattering 1001

Chapter 11

VECTORS AND CURVES

11-1. Introduction.

Given the curve $y = x^2$, we can easily show that the point A with coordinates (0,0) is an absolute minimum. We can also show that the area between the curve and the line $y = x + 2$ is $\frac{9}{2}$ (Figure 11-1a).

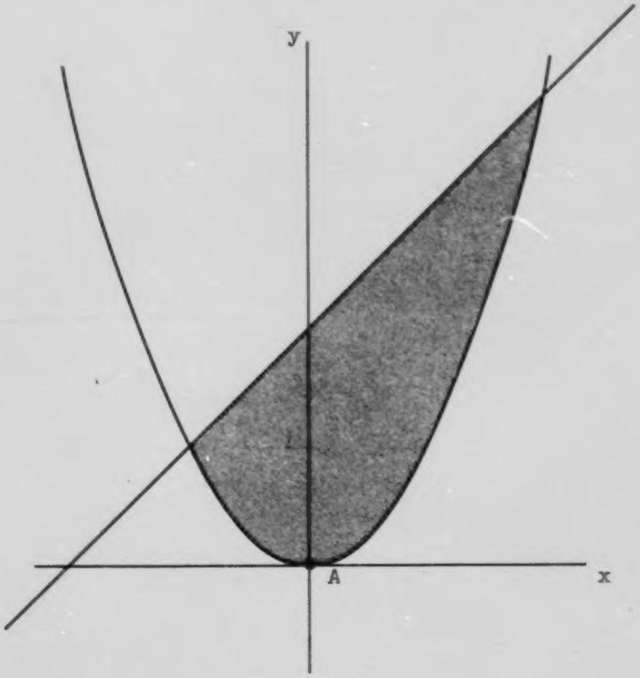


Figure 11-1a

But suppose we take the figure and rotate it about the point A until the line $y = x + 2$ is horizontal (Figure 11-1b). As a result of the change the point A is no longer a minimum. On the other hand, the area of the shaded region certainly should not change.

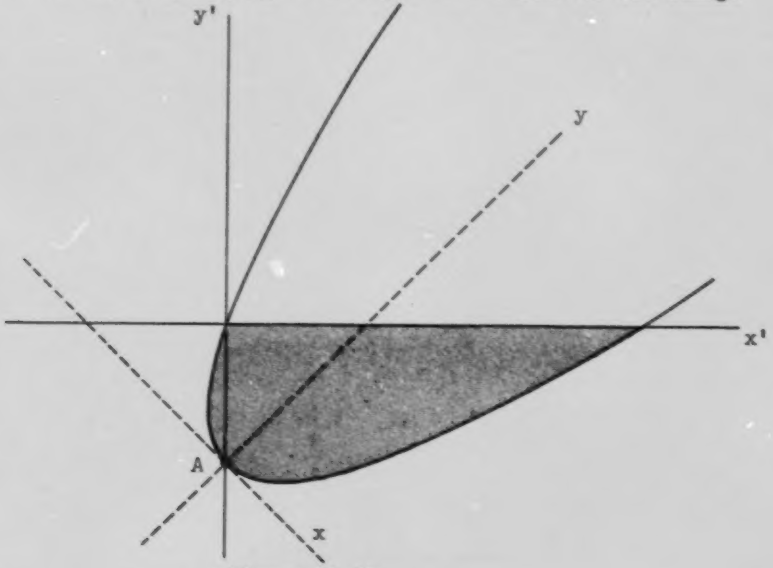


Figure 11-1b

This simple example suggests that there are two types of properties associated with curves in the plane or in space. Certain properties depend on the orientation of the curve with respect to a set of coordinate axes, and others do not. In this chapter and the next we shall emphasize those properties which do not depend on the coordinates. You should not suppose that either type of property is more important or more fundamental than the other, for they have different uses. In Section 1-1 to find a box with the largest volume, subject to given conditions, we plotted the volume of the box as a function of the length of a side (Figure 1-1a). If we were to rotate the reference axes, then the new curve would be of little use in solving the problem. In contrast, areas and volumes of regions do not depend on the coordinates. For some problems coordinates are not useful until the solution is almost complete; such a problem is that of describing the motion of two mutually attracting objects.

We shall introduce new tools particularly appropriate to the study of properties independent of coordinate axes. We combine parts of euclidean synthetic geometry and the analytic geometry of Descartes. Synthetic geometry does not employ coordinate axes but is very awkward for quantitative studies. Cartesian geometry is an ideal tool for the definition and analysis of curves and surfaces, but the analysis rests on the choice of coordinates, even when the results do not. To combine the advantages of both systems we introduce the concept of vector as a translation of the plane or space, although later the concept will be seen to have many other interpretations. We shall develop an algebra of vectors with operations corresponding to simple geometric constructions. In this way, we can describe space independently of coordinates and yet have the power of algebraic operations. All of this could have been done before the calculus, but we finally introduce a vector calculus, and bring to bear the tools we have developed in earlier chapters.

Exercises 11-1

1. Which of the following quantities are independent of the choice of coordinates?

   (a) The distance between two points.
   (b) The distance of a point from the origin.
   (c) The angle between two lines.
   (d) The angle of inclination of a line.
   (e) The area of a standard region under the graph $y = f(x)$.
   (f) The area bounded by two curves.

11-2. Vector Algebra.

We shall use the geometrical idea of translation as our primary representation of the concept of vector. A vector $\vec{V}$ is then thought of as a mapping of space[*] $\vec{V} : P \longrightarrow Q$ onto itself in which all points are translated the same distance (the length of $\vec{V}$, written $|V|$) in the same direction. Thus if $\vec{V}(P_1) = Q_1$ and $\vec{V}(P_2) = Q_2$ then the directed line segments $\overrightarrow{P_1Q_1}$ and $\overrightarrow{P_2Q_2}$ have the same length and are parallel (directed line segments are said to be parallel if they not only lie on parallel lines, but also have the same orientation). We can describe $\vec{V}$ by the way in which it maps just one point since any directed segment $\overrightarrow{PQ}$, where $Q = \vec{V}(P)$, defines the direction and length of $\vec{V}$. It does no harm to picture $\vec{V}$ as a floating directed segment (of a specified length and direction) which may be attached to an initial point if convenient.

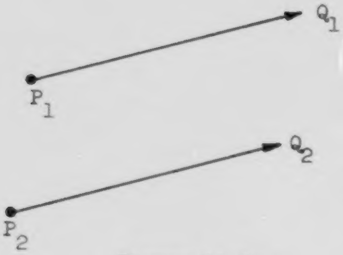


Figure 11-2a

The interpretation of vector as a translation clearly makes no use of a coordinate system. Nonetheless, it is often convenient to have a representation of a vector in a given coordinate system. If $\vec{V}(P_1) = Q_1$ and

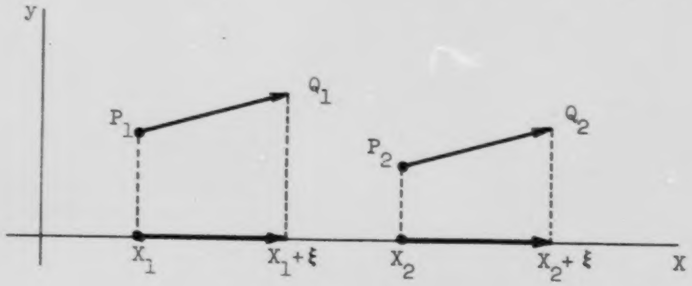


Figure 11-2b

---

[*] We shall use the word "space" to denote either the Euclidean plane E^2 or Euclidean three-dimensional space E^3, specifying only when necessary.

$\vec{V}(P_2) = Q_2$ then from the parallelism of $\overrightarrow{P_1Q_1}$ and $\overrightarrow{P_2Q_2}$ the projections of the two directed segments on any given line are equal in length and have the same orientation. Thus if P_1 and P_2 have the x-coordinates x_1 and x_2 respectively, then Q_1 and Q_2 have the x-coordinates $x_1 + \xi$ and $x_2 + \xi$, respectively (Figure 1-2b). Here $|\xi|$ is the common length of the two projections and ξ is positive if the projections have the same orientation as the x-axis, and negative if they have the opposite orientation. The number ξ is called the x-component V_x of $\vec{V}$. In exactly the same way, we introduce the y-component V_y by means of projections on the y-axis, and the z-component V_z by means of projections on the z-axis. If we happen to be concerned about the Euclidean plane E^2, not three-dimensional space E^3, then, of course, there is no z-axis and no z-component. It is convenient, however, to consider E^2 as part of E^3 by taking E^2 as the plane $z = 0$. The translations in E^2 are then simply the vectors $\vec{V}$ with $V_z = 0$. By means of this imbedding of E^2 in E^3 the statements we make about E^3 can be specialized to yield statements about E^2.

From the definition of component, we see that the effect of $\vec{V}$ upon the coordinate representation (a,b,c) of a point is to add its components to the corresponding coordinates; namely

$$(1) \qquad \vec{V} : (a,b,c) \longrightarrow (a + V_x, b + V_y, c + V_z) .$$

Thus the mapping $\vec{V}$ is completely described in a given coordinate frame by the specification of its components. We write the coordinate representation of a vector as the ordered triple of components

$$(2) \qquad \vec{V} = (V_x, V_y, V_z) .$$

In (2) we have deliberately adopted the same notation as that for the coordinate representation of a point; we shall see that this is a convenience rather than a cause for confusion. In particular, the vector $\vec{V}$ maps the origin onto a point whose coordinates are the components of $\vec{V}$,

$$(3) \qquad \vec{V} : (0,0,0) \longrightarrow (V_x, V_y, V_z)$$

thus the length of $\vec{V}$ is the distance between the points $(0,0,0)$ and (V_x, V_y, V_z), namely,

$$(4) \qquad |\vec{V}| = \sqrt{V_x^2 + V_y^2 + V_z^2} .$$

We may represent the translation $\vec{V}$ by the point into which it maps the origin, and given either one, the point or the vector, the other is determined. Thus, given an origin, there is a one-to-one correspondence between points and vectors. (This depends on the choice of origin but not the orientation of the axes, although the coordinates of points and vectors do depend on orientation.) As we shall soon see, this correspondence is given a special significance in terms of the operation of composition of translations. When we wish to emphasize the distinction we shall write $V = (V_x, V_y, V_z)$ for the point and $\vec{V} = (V_x, V_y, V_z)$ for the vector.

The composition of two translations, first $\vec{V}$ then $\vec{U}$ applied to a point P has a simple geometrical interpretation. Let $Q = \vec{V}(P)$ and $R = \vec{U}(Q)$, (Figure 1-2c).

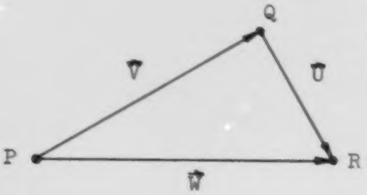


Figure 11-2c

Think of $\vec{V}$ as attached to the initial point P to yield the terminal point Q and $\vec{U}$ as attached to the tip of the arrow $\overrightarrow{PQ}$ to yield the final point R. The composition $\vec{W} = \vec{U}\vec{V}$ is the translation defined by the directed segment $\overrightarrow{PR}$. Thus we may think of $\vec{W}$ as forming the third side of a triangle for which $\vec{U}$ and $\vec{V}$ define the first two sides (provided that P, Q, R are not collinear, of course). It is natural to compare the composition $\vec{W} = \vec{U}\vec{V}$ with the composition in the reverse order, $\vec{W}^* = \vec{V}\vec{U}$ applied to the same point P. For this purpose set $S = \vec{U}(P)$ and $R^* = \vec{V}(S) = \vec{W}^*(P)$. We shall prove

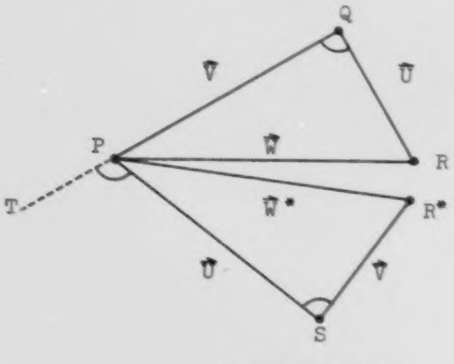


Figure 11-2d

that $R^* = R$, in other words, that the composition of translations is commutative. We leave as an exercise the special case in which $\vec{U}$ and $\vec{V}$ are parallel. First observe in Figure 11-2d that since $\overline{PQ}$ and $\overline{SR^*}$ are parallel that the exterior angle, $\angle TPS$, is equal to $\angle PSR^*$, and since $\overline{PS}$ and $\overline{QR}$ are parallel, that $\angle TPS = \angle PQR$. Thus $\angle PQR = \angle PSR^*$; hence, from the equality of the sides of the angle, $\triangle PQR$ is congruent to $\triangle R^*SP$. Consequently, $|\vec{W}^*| = |\vec{W}|$. Now, from the parallelism of $\overline{PS}$ and $\overline{QR}$ the four points P, Q, S, R are coplanar, and from the parallelism of $\overline{PQ}$ and $\overline{SR^*}$ the points P, Q, S, R^* are coplanar. Hence all five points P, Q, S, R, R^* are coplanar and Figure 11-2d represents a plane figure. But now observe that

$$\angle R^*PS = \angle PRQ = \angle RPS$$

since $\overline{QR}$ and $\overline{PS}$ are parallel. Consequently, R^* and R lie in the same direction from P at the same distance $(|\vec{W}| = |\vec{W}^*|)$, hence $R = R^*$.

The same result can be obtained very simply by means of the coordinate representations of $\vec{U}$ and $\vec{V}$. Set $P = (a,b,c)$, $\vec{U} = (U_x,U_y,U_z)$, $\vec{V} = (V_x,V_y,V_z)$. Then, by (1),

$$\vec{U}\vec{V}(P) = (a + V_x + U_x, b + V_y + U_y, c + V_z + U_z) .$$

Since the components of the composition are simply the sums of the corresponding components of the constituent translations, the composition $\vec{U}\vec{V}$ is naturally called the sum of $\vec{U}$ and $\vec{V}$ and is written $\vec{U} + \vec{V}$. For the coordinate representation of the sum, we have

$$(5) \qquad \vec{U} + \vec{V} = (U_x + V_x, U_y + V_y, U_z + V_z) .$$

Since addition is commutative, it follows that $\vec{U} + \vec{V} = \vec{V} + \vec{U}$. The commutativity of the sum leads to the "parallelogram" law for the addition of vectors. In Figure 11-2d the points R and R^* should be the same and the figure $PQRS$ is a parallelogram for which the directed diagonal $\overrightarrow{PR}$ represents $\vec{U} + \vec{V}$.

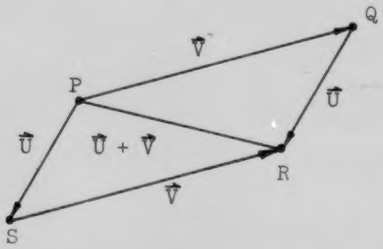


Figure 11-2e

From (5) we conclude that addition of vectors is also associative:

(6)
$$(\vec{U} + \vec{V}) + \vec{W} = \vec{U} + (\vec{V} + \vec{W}) .$$

The geometrical proof of the associative law (6) is left to Exercises 11-2, Number 6. It is almost as obvious as the algebraic proof.

Formulas (1) and (5), the one representing the effect of a translation, the other, the addition of two vectors, are formally identical. As far as the algebraic consequences are involved, it does not matter whether we distinguish the point $P = (a,b,c)$ from the "position" vector $\vec{P} = (a,b,c)$ which maps the origin onto P. Once we have chosen an origin we shall feel free henceforth to use point and vector as interchangeable terms.

In order to complete the analogy of vector addition with ordinary addition we introduce the null vector $\vec{0}$ which may be thought of as the "identity" translation which maps each point onto itself. Thus $\vec{0}$ has the coordinate representation

(7)
$$\vec{0} = (0,0,0)$$

and satisfies the property

(8)
$$\vec{0} + \vec{V} = \vec{V} + \vec{0} = \vec{V}$$

for all vectors $\vec{V}$.

Each translation $\vec{V}$ has an (additive) inverse $-\vec{V}$ which undoes the effect of $\vec{V}$. Thus, if $\vec{V}(P) = Q$ then $-\vec{V}(Q) = P$. Consequently, $-\vec{V}$ is a vector of the same length as $\vec{V}$, but opposite direction. The inverse $-\vec{V}$ satisfies

(9)
$$\vec{V} + (-\vec{V}) = \vec{0} ;$$

hence, if $\vec{V} = (V_x,V_y,V_z)$ the coordinate representation of $-\vec{V}$ is given by

(10)
$$-\vec{V} = (-V_x,-V_y,-V_z) .$$

A second useful operation is multiplication of a vector by a real number. By the vector $\lambda\vec{V}$, where $\lambda > 0$, is meant the vector in the direction of $\vec{V}$ with length λ times that of $\vec{V}$. If $\lambda < 0$, then $\lambda\vec{V}$ is taken with length $-\lambda$ times the length of $\vec{V}$ and direction opposite to that of $\vec{V}$. If $\lambda = 0$ then we take $\lambda\vec{V} = 0$. In any case,

(11)
$$|\lambda\vec{V}| = |\lambda| \cdot |\vec{V}| .$$

Note further that

(12)
$$(-\lambda)\vec{V} = -(\lambda\vec{V}) .$$

If we interpret $\vec{V}$ as a point we see for $\lambda > 0$ that the operation of multiplying λ applied to all points of space amounts to a change of scale or magnification by the factor λ. For this reason, we call real numbers scalars. Finally, in terms of the coordinate representation of V,

$$\lambda\vec{V} = (\lambda V_x, \lambda V_y, \lambda V_z) \tag{13}$$

which means geometrically that to magnify $\vec{V}$ by a factor λ is equivalent to magnifying each component of $\vec{V}$ by the same factor. From (13) it follows directly that multiplication by a scalar is associative,

$$\lambda(\mu\vec{V}) = (\lambda\mu)\vec{V}, \tag{14}$$

and satisfies the distributive laws

$$(\lambda + \mu)\vec{V} = \lambda\vec{V} + \mu\vec{V} \tag{15}$$

$$\lambda(\vec{U} + \vec{V}) = \lambda\vec{U} + \lambda\vec{V}. \tag{16}$$

The usual connections between the concepts of subtraction, inverse, and multiplication by -1 still hold. Thus we define the difference $\vec{V} - \vec{U}$ by

$$\vec{V} - \vec{U} = \vec{V} + (-\vec{U}).$$

Geometrically $\vec{V} - \vec{U}$ corresponds to the second diagonal of the parallelogram in the parallelogram law for $\vec{V} + \vec{U}$ (i.e., the segment SQ in Figure 11-2e); the verification is left to Exercises 11-2, Number 2. From their geometrical definitions it is immediate that $-\vec{V} = (-1)\vec{V}$.

If $\vec{V} = \mu\vec{U}$ or $\vec{U} = \lambda\vec{V}$ we say that the vectors $\vec{U}$ and $\vec{V}$ are collinear. This corresponds to the conventional statement that the points O, U, and V are collinear. Note that the vector $\vec{0}$ is by this definition collinear with every vector. (See Exercises 11-2, No. 4(d))

In the preceding discussion we have been guided by geometrical ideas, but there are many other ways of representing vectors than the ones we have chosen. The basic algebraic structure we have exhibited is the defining characteristic of vectors. In general a set $\mathcal{L}$ is called a linear vector space over the real numbers and its elements called vectors if there are two operations, addition of vectors and multiplication of vectors by a scalar, which obey the following laws.

For each pair of vectors $\vec{U}$ and $\vec{V}$ in $\mathcal{L}$ there is a vector $\vec{U} + \vec{V}$ in $\mathcal{L}$, called the sum of $\vec{U}$ and $\vec{V}$ with the following properties.[1]

A1. Commutativity, $\vec{U} + \vec{V} = \vec{V} + \vec{U}$.

A2. Associativity, $(\vec{U} + \vec{V}) + \vec{W} = \vec{U} + (\vec{V} + \vec{W})$.

A3. There exists a vector $\vec{0}$ in $\mathcal{L}$, called the null vector such that $\vec{V} + \vec{0} = \vec{V}$ for each vector $\vec{V}$ in $\mathcal{L}$.

A4. For each vector $\vec{V}$ in $\mathcal{L}$ there exists a vector $-\vec{V}$ such that $\vec{V} + (-\vec{V}) = \vec{0}$.

For each scalar (real number) λ and each vector $\vec{V}$ in $\mathcal{L}$ there is a vector $\lambda\vec{V}$ in $\mathcal{L}$, called the product of λ and $\vec{V}$, with the following properties.[2]

M1. $1\vec{V} = \vec{V}$.

M2. $\lambda(\mu\vec{V}) = (\lambda\mu)\vec{V}$.

Multiplication by vectors is distributive over addition of scalars,

D1. $(\lambda + \mu)\vec{V} = \lambda\vec{V} + \mu\vec{V}$.

Multiplication by scalars is distributive over addition of vectors,

D2. $\lambda(\vec{U} + \vec{V}) = \lambda\vec{U} + \lambda\vec{V}$.

The postulates for a linear vector space are given above only to present the abstract mathematical concept of vector precisely. It includes spaces which you might not at first connect with our geometrical model. For example, the set of all polynomials with real coefficients form a linear vector space; the set of all solutions of a linear homogeneous differential equation is another. For our present purposes, however, the geometrical model is sufficient. We leave to the exercises some of the simpler algebraic consequences of the vector space postulates.

---

[1] The Properties A1 - 4 define a structure, called an abelian or commutative group, which appears in many different contexts.

[2] We adhere to the convention that the product of a scalar and a vector is always written with the scalar on the left.

Exercises 11-2

1. Let U and V be any points and $\vec{U}$, $\vec{V}$ the corresponding position vectors. In terms of $\vec{U}$ and $\vec{V}$ what vector is represented by the directed segment $\overrightarrow{UV}$?

2. In Figure 11-2e, one diagonal corresponds to the sum $\vec{U} + \vec{V}$. What vector corresponds to the other diagonal?

3. Give a geometrical justification for the inequality

$$|\vec{U} + \vec{V}| \leq |\vec{U}| + |\vec{V}| .$$

4. Let A and B be any given points. Characterize geometrically each of the sets of points

   (a) $\{X : |\vec{X} - \vec{A}| = r\}$.

   (b) $\{X : |\vec{X} - \vec{A}| < r\}$.

   (c) $\{X : |\vec{X} - \vec{A}| > r\}$.

   (d) $\{X : \vec{X} = \lambda\vec{A} , \lambda \text{ real}\}$.

   (e) $\{X : \vec{X} = \lambda\vec{A} , \lambda \geq 0\}$.

   (f) $\{X : \vec{X} = \vec{A} + \lambda\vec{B} , \lambda \geq 0\}$.

   (g) $\{X : \vec{X} = \vec{A} + \lambda\vec{B} , \lambda \text{ real}\}$.

   (h) $\{X : |\vec{X} - \vec{A}| = |\vec{X} - \vec{B}|\}$.

5. For any non-null vector $\vec{A}$ obtain the unit vector (vector of length 1) in the direction of $\vec{A}$.

6. Give a geometrical derivation for the associative law (6) for the addition of vectors.

7. From the laws of operation, A1 - 4, M1 - 2, D1 - 2, which define a vector space, derive the following consequences.

   (a) $\lambda\vec{0} = \vec{0}$

   (b) $0\vec{V} = \vec{0}$

   (c) If $\lambda \neq 0$ and $\vec{V} \neq \vec{0}$ then $\lambda\vec{V} \neq \vec{0}$.

   (d) $(-1)\vec{V} = -\vec{V}$

   (e) If $\lambda \neq 0$, the vector equation $\lambda\vec{X} + \vec{U} = \vec{V}$ has the unique solution $\vec{X} = \frac{1}{\lambda}(\vec{V} - \vec{U})$.

8. Show that the set of continuous functions on the interval $[0,1]$ is a linear vector space over the real numbers where addition and multiplication have their conventional interpretations.

9. (a) Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be linear vector spaces over the real numbers. Show that the set of ordered pairs $\mathcal{L}_1 \oplus \mathcal{L}_2 = \{(\vec{V}_1,\vec{V}_2) : \vec{V}_1 \epsilon \mathcal{L}_1, \vec{V}_2 \epsilon \mathcal{L}_2\}$ is a linear vector space over the real numbers, where, for $\vec{U}_1$, $\vec{V}_1 \epsilon \mathcal{L}_1$ and $\vec{U}_2$, $\vec{V}_2 \epsilon \mathcal{L}_2$ addition and multiplication by a scalar in $\mathcal{L}_1 \oplus \mathcal{L}_2$ are defined by

$$(\vec{U}_1,\vec{U}_2) + (\vec{V}_1,\vec{V}_2) = (\vec{U}_1 + \vec{V}_1 , \vec{U}_2 + \vec{V}_2)$$

and

$$\lambda(\vec{V}_1,\vec{V}_2) = (\lambda\vec{V}_1,\lambda\vec{V}_2) .$$

The space $\mathcal{L}_1 \oplus \mathcal{L}_2$ is known as the direct sum of $\mathcal{L}_1$ and $\mathcal{L}_2$.

(b) Show that $\mathcal{R}$, the real number field is a linear vector space over the real numbers, where addition and multiplication is now ordinary addition and multiplication of numbers. The set $\mathcal{R}$ considered as a linear vector space with a length defined as $|x|$ for $x \epsilon \mathcal{R}$ is denoted by E^1 (one-dimensional euclidean space).

(c) Show that euclidean two-dimensional space E^2 is given by

$$E^2 = E^1 \oplus E^1 ,$$

where length for $\vec{A} \epsilon E^2$, given that $\vec{A} = (a,b)$ and $a , b \epsilon E^1$, is defined by

$$|\vec{A}| = \sqrt{a^2 + b^2} .$$

Similarly, show that euclidean three-dimensional space E^3 is given by

$$E^3 = E^2 \oplus E^1 ,$$

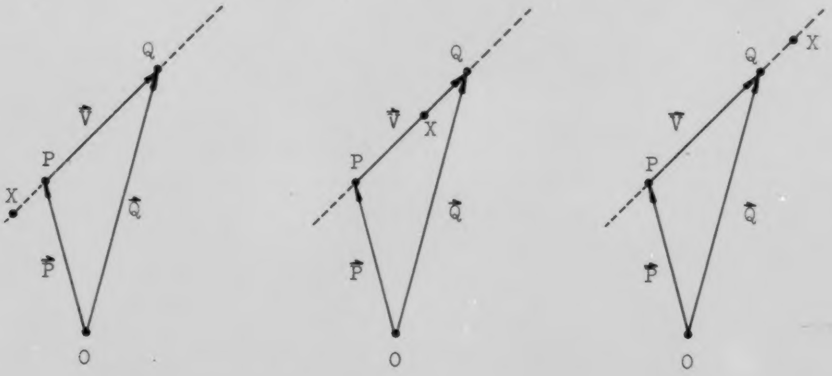
where length for $\vec{V} \epsilon E^3$, given in the form $\vec{V} = (\vec{A},c)$ with $\vec{A} \epsilon E^2$, $c \epsilon E^1$, is defined by

$$|\vec{V}| = \sqrt{|\vec{A}|^2 + c^2} .$$

11-3. Vector Geometry.

In this section, we give some examples of geometrical theorems using vector methods. In many cases, the proofs are much simpler than the proofs of synthetic or analytic geometry.

First we shall establish some preliminaries. We choose an origin 0 so that we may represent a point as a position vector. Let P and Q be any



$\lambda < 0$
(i)

$0 < \lambda < 1$
(ii)

$\lambda > 1$
(iii)

Figure 11-3a

two points and $\vec{P}$ and $\vec{Q}$ the corresponding position vectors (Figure 11-3a). The line PQ is parallel to the vector $\vec{V}$ represented by the directed segment $\overrightarrow{PQ}$. Thus a point X of the line may be represented as the result of a translation from 0 to P followed by a translation in the direction of $\vec{V}$ or its inverse, and we may write $\vec{X}$ in the form $\vec{X} = \vec{P} + \lambda\vec{V}$ where the parameter may be any real number. In particular, we observe that if λ increases from 0 to 1, X traverses the segment $\overline{PQ}$ from P to Q and that λ represents the ratio of the length of $\overline{PX}$ to that of $\overline{PQ}$, (Figure 11-3a(ii)). If $\lambda > 1$, then X lies on the other side of Q from P, (Figure 11-3a(iii)), and if $\lambda < 0$ then X lies on the other side of P from Q, (Figure 11-3a(i)). We observe also that the vector $\vec{V}$ represented by $\overrightarrow{PQ}$ satisfies $\vec{P} + \vec{V} = \vec{Q}$; hence, $\vec{V} = \vec{Q} - \vec{P}$. Thus we obtain the vector form for the two-point equation of the line,

$$\vec{X} = (1 - \lambda)\vec{P} + \lambda\vec{Q} . \tag{1}$$

Here $|\lambda|$ has the geometrical meaning of the ratio of the length of $\overline{PX}$ to that of $\overline{PQ}$, λ being positive if Q and X lie in the same direction from

P , negative if Q and X lie in opposite directions from P . Hence, in particular, the midpoint of the segment $\overline{PQ}$ is given by $\lambda = \frac{1}{2}$:

$$\vec{X} = \frac{1}{2}(\vec{P} + \vec{Q}) \quad (2)$$

To gain some experience in vector manipulations we now use the vector approach to prove some geometrical theorems.

Example 11-3a. The midpoints of the sides of a quadrilateral are the vertices of a parallelogram.

Let the vertices of the quadrilateral be A , B , C , D , and the corresponding position vectors, $\vec{A}$, $\vec{B}$, $\vec{C}$, $\vec{D}$ (see Figure 11-3a).

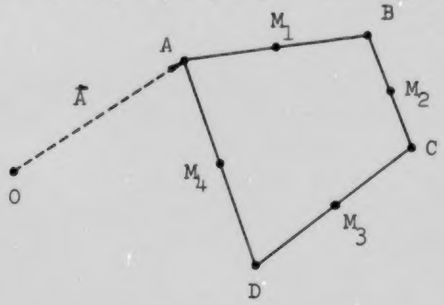


Figure 11-3a

The midpoint M_1 of $\overline{AB}$ is given by

$$\vec{M}_1 = \frac{1}{2}(\vec{A} + \vec{B}) .$$

Similarly, the midpoints of the other sides are given by

$$\vec{M}_2 = \frac{1}{2}(\vec{B} + \vec{C}) , \vec{M}_3 = \frac{1}{2}(\vec{C} + \vec{D}) , \vec{M}_4 = \frac{1}{2}(\vec{D} + \vec{A}) .$$

The figure $M_1M_2M_3M_4$ is a parallelogram if and only if $\overrightarrow{M_1M_4}$ and $\overrightarrow{M_2M_3}$ are parallel and equal in length; i.e., if and only if $\vec{M}_4 - \vec{M}_1 = \vec{M}_3 - \vec{M}_2$. But $\vec{M}_4 - \vec{M}_1 = \frac{1}{2}(\vec{D} - \vec{B}) = \vec{M}_3 - \vec{M}_2$, which proves the theorem.

Note that the proof holds even if the original quadrilateral does not lie in a plane.

Example 11-3b. The arithmetic average A of n points P_1 , P_2 , ... , P_n is given by $\vec{A} = \frac{1}{n}(\vec{P}_1 + \vec{P}_2 + \ldots + \vec{P}_n)$ and is called their centroid.

The three medians of a triangle are concurrent at the centroid of the vertices. Furthermore the distance from the centroid to the midpoint of any side is half that to the opposite vertex.

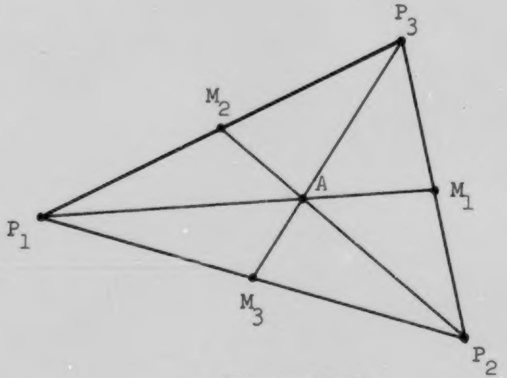


Figure 11-3b

Let P_1 , P_2 , P_3 be the vertices of the triangle and M_1 , M_2 , M_3 the midpoints of the respective opposite sides. The equation of the median from P_1 to M_1 is

$$\begin{aligned}\vec{X} &= (1 - \lambda)\vec{P}_1 + \lambda\vec{M}_1 \\ &= (1 - \lambda)\vec{P}_1 + \frac{\lambda}{2}(\vec{P}_2 + \vec{P}_3) \ .\end{aligned}$$

The centroid A is on this line if and only if there is a value of λ for which X = A , i.e.,

$$\frac{\vec{P}_1 + \vec{P}_2 + \vec{P}_3}{3} = (1 - \lambda)\vec{P}_1 + \frac{\lambda}{2}\vec{P}_2 + \frac{\lambda}{2}\vec{P}_3 \ ,$$

or

$$(\frac{2}{3} - \lambda)\vec{P}_1 + (\frac{\lambda}{2} - \frac{1}{3})\vec{P}_2 + (\frac{\lambda}{2} - \frac{1}{3})\vec{P}_3 = \vec{0} \ .$$

This last equation is clearly satisfied by $\lambda = \frac{2}{3}$. Thus we see that the centroid subdivides the median in the stated ratio. The same argument applied to the other medians completes the proof.

Just as a line may be described by any two of its points, a plane may be described by any noncollinear triple of its points. Let A , B , C be any noncollinear triple of points. The plane ABC must contain with any two of its points the entire line on which they lie. It follows, then, that the plane contains the two lines $\{P : \vec{P} = \vec{A} + \mu(\vec{B} - \vec{A}) \ , \mu \text{ real}\}$ and

$\{Q : \vec{Q} = \vec{A} + \nu(\vec{C} - \vec{A})\ ,\ \nu \text{ real}\}$, consequently, it contains the set of points

$$(3)\quad \{R : \vec{R} = (1 - \lambda)\vec{P} + \lambda\vec{Q} = \vec{A} + (1 - \lambda)\mu(\vec{B} - \vec{A}) + \lambda\nu(\vec{C} - \vec{A})\ ,\ \lambda\ ,\ \mu\ ,\ \nu \text{ real}\}\ .$$

Furthermore, any point of the plane is a point of this set; for if R is a point of the plane, take any line through R which is not parallel to either

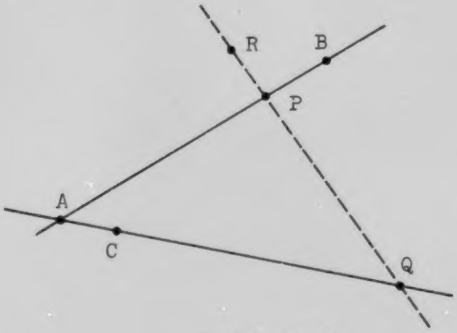


Figure 11-3c

the line AB or the line AC and does not pass through A . The line through R will meet AB at a point P and AC at a second point Q . It follows that P , Q and R can be written in the forms given above.

The form (3) in which we have represented the plane ABC is unnecessarily complex. For simplicity, set $(1 - \lambda)\mu = \xi$ and $\lambda\nu = \eta$. Observe that any pair of real numbers ξ , η can be represented in this way by taking $\lambda = \frac{1}{2}$, $\mu = 2\xi$, $\nu = 2\eta$. Consequently, we may put

$$(4)\qquad ABC = \{R : \vec{R} = \vec{A} + \mu(\vec{B} - \vec{A}) + \nu(\vec{C} - \vec{A})\ ,\ \mu\ ,\ \nu \text{ real}\}\ .$$

The representation (4) of ABC has a simple geometrical interpretation. Take A as the origin of a coordinate system in the plane with ξ , η-axes in the directions of the vectors $\vec{U} = \vec{B} - \vec{A}$ and $\vec{V} = \vec{C} - \vec{A}$, respectively. (Such axes will usually not be perpendicular.) Choose scales (generally unequal) along the axes with $|\vec{U}|$ as the unit along the ξ-axis and $|\vec{V}|$ as the unit along the η-axis. A point R of the plane is given by the coordinates (α,β) , the position of R being determined by a translation $\alpha\vec{U}$ of A along the ξ-axis, followed by a translation $\beta\vec{V}$ parallel to the η-axis (see Figure 11-3d).

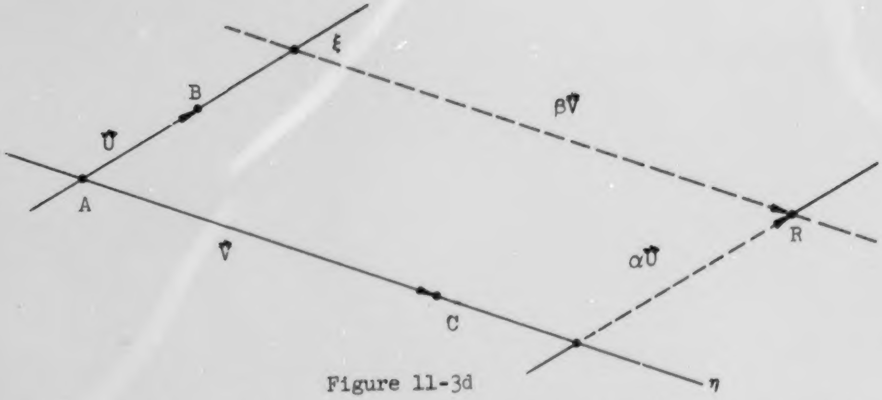


Figure 11-3d

This procedure is a straightforward generalization of the familiar method for locating a point by its rectangular coordinates.

If the points O , A , B , C are coplanar we say also that the vectors $\vec{A}$, $\vec{B}$, $\vec{C}$ are coplanar. Note that by this definition any three vectors which include the null vector are coplanar.

Example 11-3c. The vectors $\vec{X}$, $\vec{Y}$, $\vec{Z}$ are coplanar if and only if there exist three scalars a , b , c , not all zero, such that

$$a\vec{X} + b\vec{Y} + c\vec{Z} = \vec{0} \; .$$

(a) Suppose $a\vec{X} + b\vec{Y} + c\vec{Z} = \vec{0}$, say $a \neq 0$, then

$$\vec{X} = -\frac{b}{a}\vec{Y} - \frac{c}{a}\vec{Z} \; .$$

By reference to (4) above with $\vec{A} = \vec{0}$, $B = Y$, and $C = Z$ we see that X is in the plane determined by Y , Z and the origin O .

(b) Suppose $\vec{X}$, $\vec{Y}$, $\vec{Z}$ are coplanar. If any one of them, say $\vec{X}$, is the null vector then we have the relation

$$1 \cdot \vec{X} + 0 \cdot \vec{Y} + 0 \cdot \vec{Z} = \vec{0} \; ,$$

so we assume none of the vectors is $\vec{0}$. If $\vec{Y}$ is collinear with $\vec{X}$ then $\vec{Y} = \lambda\vec{X}$, and

$$\lambda\vec{X} - \vec{Y} + 0 \cdot \vec{Z} = \vec{0} \; .$$

Finally, if O , X , and Y are not collinear, then they determine a plane, and by (4), any Z which lies in the plane satisfies

$$\vec{Z} = \vec{0} + \mu\vec{X} + \nu\vec{Y}$$

or

$$\mu\vec{X} + \nu\vec{Y} - \vec{Z} = \vec{0} \; .$$

Corollary. Suppose $\vec{X}$, and $\vec{Y}$ are not collinear, then any point Z in the plane OXY has the unique representation

$$\vec{Z} = a\vec{X} + b\vec{Y} .$$

We leave to Exercise 11-3, Number 8 the proof that given any four vectors $\vec{W}$, $\vec{X}$, $\vec{Y}$, $\vec{Z}$ in E^3 there are constants a , b , c , d not all zero so that

$$a\vec{W} + b\vec{X} + c\vec{Y} + d\vec{Z} = \vec{0} ,$$

and hence given any three vectors $\vec{X}$, $\vec{Y}$, $\vec{Z}$, not coplanar, any vector $\vec{W}$ has the representation $\vec{W} = \alpha\vec{X} + \beta\vec{Y} + \gamma\vec{Z}$.

We introduced position vectors referred to a given origin O . We now consider how a position vector is affected by a shift of origin to another point A . Let $\vec{X}$ be any position vector referred to O . The vector $\vec{X} - \vec{A}$ is represented by the directed segment $\overrightarrow{AX}$ and therefore is the position vector of X referred to A as origin (Figure 11-3e). Thus, instead of shifting the origin to A , we may translate the plane by $-\vec{A}$ and the result is algebraically the same. Note that any result which is unaffected by the transformation $\vec{X} \longrightarrow \vec{X} - \vec{A}$ is independent of the choice of origin (see Exercises 11-3, No. 6).

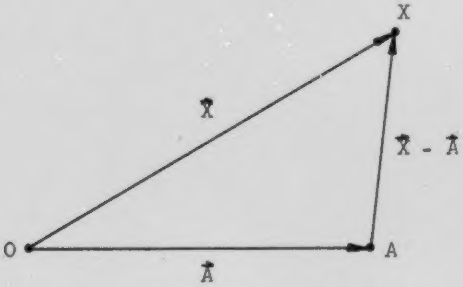


Figure 11-3e

Exercises 11-3

1. Verify that Formula (2) gives the familiar formulas for the midpoint of a segment in terms of coordinates for P and Q.

2. Find the equation of the line through the point $P = (1,2,1)$ parallel to the position vector $(0,3,4)$; then give the coordinates of a point on the line at distance 1 from P.

3. (a) Is $P = (2,1,2)$ in the plane P containing $A = (1,1,3)$, $B = (1,1,2)$, $C = (1,3,3)$?

   (b) Find a vector $\vec{N}$ normal to the plane P.

   (c) Find the distance from $P = (2,1,2)$ to the plane P.

4. Prove the corollary in Example 11-3c.

5. Let $\vec{A}$ and $\vec{B}$ be noncollinear vectors. Determine the equation of the ray which bisects $\angle AOB$.

6. Verify that the results of (1), of Examples 11-3a, b, and of (4) do not depend on the choice of origin.

7. (a) Show that the vectors $\vec{A} = (1,1,3)$, $\vec{B} = (1,1,2)$ and $\vec{C} = (1,3,3)$ are noncoplanar.

   (b) Express the vector $\vec{D} = (2,1,2)$ in the form of a linear combination

$$\vec{D} = a\vec{A} + b\vec{B} + c\vec{C}.$$

8. Show that given any four vectors $\vec{A}$, $\vec{B}$, $\vec{C}$, $\vec{D}$, there are constants a, b, c, d, not all zero, so that

$$a\vec{A} + b\vec{B} + c\vec{C} + d\vec{D} = \vec{0}.$$

   (Hint: Use the property that if a line is not parallel to a plane it must intersect that plane at precisely one point.)

9. The statement of Number 8 implies that any four vectors are linearly dependent, namely that one can be expressed as a linear combination of the other three. Show from the assumption that E^3 is not contained in a plane that there do exist three linearly independent vectors in E^3, that is, vectors $\vec{A}$, $\vec{B}$, $\vec{C}$ for which the equation

$$a\vec{A} + b\vec{B} + c\vec{C} = \vec{0}$$

   is satisfied only if all three scalars a, b, c are zero.

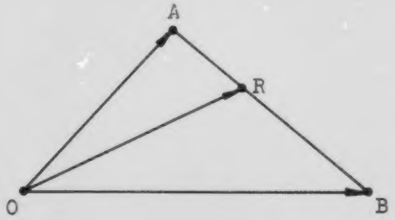
10. Prove if $\vec{A}$, $\vec{B}$, $\vec{C}$ are not coplanar then any vector $\vec{Z}$ can be represented as a linear combination

$$\vec{Z} = a\vec{A} + b\vec{B} + c\vec{C}$$

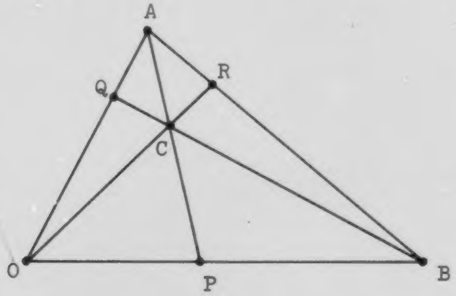
and that the representation is unique.

11. Show if $\vec{B}$ and $\vec{C}$ are not collinear that $\vec{X} = \vec{A} + p\vec{B} + q\vec{C}$ is the equation of a plane passing through the point $\vec{A}$ and parallel to the vectors $\vec{B}$ and $\vec{C}$. We say that a plane is parallel to a vector $\vec{V}$ if it contains a directed segment $\overrightarrow{PQ}$ which represents $\vec{V}$.

12. (a) In the accompanying figure, R is any point on the line AB . Obtain the representation $\vec{R} = a\vec{A} + b\vec{B}$, and determine $\frac{|\overline{AR}|}{|\overline{RB}|}$ in terms of a and b .



∧ (b) In the accompanying figure C is any point not on a side of the triangle.



Set $\vec{R} - \vec{A} = \alpha(\vec{B} - \vec{R})$, $\vec{P} - \vec{B} = \beta(\vec{O} - \vec{P})$, $\vec{Q} - \vec{O} = \gamma(\vec{A} - \vec{Q})$. Show that

$$\alpha\beta\gamma = 1 .$$

This result together with its converse (namely, if P , Q , R divide their respective sides so that this relation holds, then the lines AP , BQ , and OR are concurrent) is Ceva's Theorem (Giovanni Ceva, Italian, 1647 - 1736).

13. Let (OABC) be a parallelogram, D the midpoint of $\overline{BC}$, and E the midpoint of $\overline{CO}$. Show that the lines AD and AE divide the diagonal $\overline{OB}$ in thirds.

Λ14. Let $P_1, P_2, \ldots, P_n$ be the consecutive vertices of a regular polygon of n sides. Show that $\vec{P}_1 + \vec{P}_2 + \ldots + \vec{P}_n = \vec{0}$ where 0 is the center of the polygon. Is this result independent of the choice of origin?

15. Prove that the bisectors of the interior angles of a triangle are concurrent.

16. A median of a tetrahedron is a line segment joining any one vertex to the centroid of the other three. Show that the medians of the tetrahedron are concurrent at the centroid of its four vertices. Show also that the segment of median between the centroid and vertex is $\frac{3}{4}$ of the total length of the median.

17. The segment joining the midpoint of any edge of a tetrahedron to the midpoint of the opposite edge is bisected by the centroid of the four vertices.

18. The eight planes, each containing one edge and bisecting the opposite edge of a tetrahedron, are concurrent.

19. Let $\vec{A}$, $\vec{B}$ be any noncollinear vectors, let $\mathcal{L}$ be any line in the plane OAB which contains no vertex of the triangle $\overline{OAB}$ and is parallel to no side. Let P be the intersection of $\mathcal{L}$ with OB, Q the intersection with OA, and R the intersection with AB. If α, β, γ are given by

$$\vec{R} - \vec{A} = \alpha(\vec{B} - \vec{R}) \ , \ \vec{B} - \vec{P} = \beta\vec{P} \ , \ Q = \gamma(\vec{Q} - \vec{A})$$

show that $\alpha\beta\gamma = -1$. Conversely, if P, Q, R are points satisfying $\alpha\beta\gamma = -1$ then they are collinear. (Meneläus's Theorem.)

11-4. Products of Two Vectors.

So far we have introduced addition of vectors and multiplication of vectors by ordinary numbers. In this section we examine two useful ways of multiplying one vector by another. The properties of these products are independent of coordinate axes. At the same time a product $A \otimes B$ of two vectors will have certain linearity properties:

$$\begin{aligned}\vec{A} \otimes (\alpha\vec{B} + \beta\vec{C}) &= \alpha\vec{A} \otimes \vec{B} + \beta\vec{A} \otimes \vec{C}\\ (\alpha\vec{B} + \beta\vec{C}) \otimes \vec{A} &= \alpha\vec{B} \otimes \vec{A} + \beta\vec{C} \otimes \vec{A} ,\end{aligned}$$

where $\otimes$ represents any such operation. The linearity properties and independent of coordinate axes characterize three types of products of two vectors. Of the three kinds of product, the two we use are the dot product (also called an inner product or scalar product) $\vec{A} \cdot \vec{B}$, which is an ordinary number, and the cross product (also called the vector product) $\vec{A} \times \vec{B}$, which is a vector. Here, we only give the products and verify that they have the desired properties.

(i) The dot product. Let $\vec{A}$ and $\vec{B}$ be two position vectors and let θ be the angle between them, where $0 \leq \theta \leq \pi$, (Figure 11-4a).

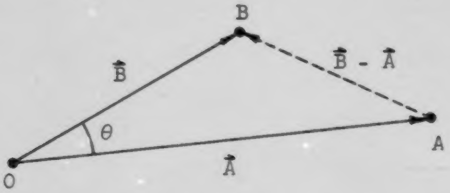


Figure 11-4a

The dot product (or scalar product) of the two vectors, written $\overline{A} \cdot \overline{B}$, is defined by

$$\vec{A} \cdot \vec{B} = |\vec{A}|\,|\vec{B}| \cos\theta . \tag{1}$$

The expression for the dot product is already familiar as a term in the law of cosines for the length of the side $\overline{AB}$ in $\triangle OAB$; namely,

$$\begin{aligned}|\vec{B} - \vec{A}|^2 &= |\vec{A}|^2 + |\vec{B}|^2 - 2|A|\,|B| \cos\theta\\ &= |\vec{A}|^2 + |\vec{B}|^2 - 2\vec{A} \cdot \vec{B} .\end{aligned} \tag{2}$$

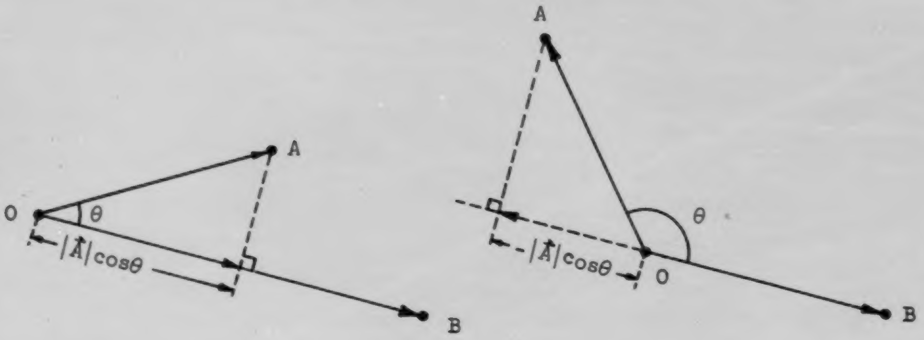
If we enter the coordinate representations $\vec{A} = (A_x, A_y, A_z)$ and $\vec{B} = (B_x, B_y, B_z)$ in (2) we obtain

$$(3) \qquad \vec{A} \cdot \vec{B} = A_x B_x + A_y B_y + A_z B_z$$

(Exercises 11-4, No. 2). However, it is clear from the Definition (1), which contains no reference to coordinates, that $\vec{A} \cdot \vec{B}$ is independent of the orientation of the coordinate axes. In (2) we have already given one invariant interpretation (that is, an interpretation independent of the orientation of the coordinate axes) of $\vec{A} \cdot \vec{B}$ but there is another which we shall also find useful. Observe that the perpendicular projection of $\vec{A}$ on the line OB has the signed length

$$(4) \qquad |\vec{A}| \cos \theta$$

where the sign is positive if the projection of A falls on the same side of O as B, $(0 \leq \theta < \frac{\pi}{2})$, and negative if it falls on the opposite side $(\frac{\pi}{2} < \theta \leq \pi)$ (see Figure 11-4b). Thus the dot product



$0 \leq \theta < \frac{\pi}{2}$ $\qquad$ $\frac{\pi}{2} < \theta \leq \pi$

Figure 11-4b

$\vec{A} \cdot \vec{B}$ is the length of $\vec{B}$ times the signed length of the projection of $\vec{A}$ on $\vec{B}$ (compare Exercises 11-4, No. 3(a)).

The dot product of two vectors vanishes if and only if one of the vectors has zero length, or if the two vectors are perpendicular $(\theta = \frac{\pi}{2})$. We adopt the convention that the null vector is both parallel and perpendicular to any vector so that $\vec{A} \cdot \vec{B} = 0$ if and only if $\vec{A}$ and $\vec{B}$ are perpendicular.

We leave as exercises the proofs of the following essential properties of the dot product (Exercises 11-4, No. 1).

(5a) $$\vec{A} \cdot \vec{B} = \vec{B} \cdot \vec{A} \, .$$

(5b) $$\vec{A} \cdot (\alpha\vec{B} + \beta\vec{C}) = \alpha\vec{A} \cdot \vec{B} + \beta\vec{A} \cdot \vec{C} \, .$$

(5c) $$|\vec{A}|^2 = \vec{A} \cdot \vec{A} \, .$$

(5d) $$|\vec{A} \cdot \vec{B}| \leq |\vec{A}|\,|\vec{B}| \, .$$

It is conventional to omit the absolute value sign in (5c) and write $\vec{A} \cdot \vec{A} = \vec{A}^2$.

As a further consequence of the properties (5), note that

$$|\vec{A} + \vec{B}|^2 = |\vec{A}|^2 + |\vec{B}|^2 + 2\vec{A} \cdot \vec{B} \leq |\vec{A}|^2 + |\vec{B}|^2 + 2|\vec{A}|\,|\vec{B}| \leq (|\vec{A}| + |\vec{B}|)^2 \, ;$$

hence,

(6) $$|\vec{A} + \vec{B}| \leq |\vec{A}| + |\vec{B}| \, .$$

This inequality justifies the use of the absolute value sign for the length of a vector. Geometrically, it states that the length of one side of a traingle is less than the sum of the lengths of the other two sides. From (6) we obtain, as for absolute values of numbers (Section A1-3), the inequality

(7) $$\Big|\, |\vec{A}| - |\vec{B}| \,\Big| \leq |\vec{A} + \vec{B}| \, .$$

The use of the relation $\vec{A} \cdot \vec{B} = 0$ for perpendicularity often simplifies the treatment of geometrical problems.

Example 11-4a. The altitudes of a triangle are concurrent.

Let the triangle have vertices determined by $\vec{A}$, $\vec{B}$, and $\vec{C}$. Let two of the altitudes intersect at $\vec{X}$, so that the vector $\vec{A} - \vec{X}$ is perpendicular

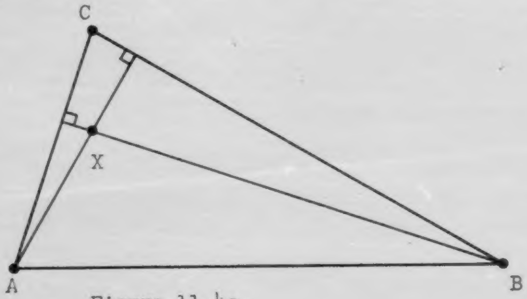


Figure 11-4c.

to $\vec{B} - \vec{C}$, and $\vec{B} - \vec{X}$ is perpendicular to $\vec{A} - \vec{C}$, (Figure 11-4c), or in terms of the dot product

$$(\vec{X} - \vec{A}) \cdot (\vec{B} - \vec{C}) = 0$$

$$(\vec{X} - \vec{B}) \cdot (\vec{C} - \vec{A}) = 0 \; .$$

We wish to show that the line through $\vec{C}$ and $\vec{X}$ is an altitude, or that $\vec{C} - \vec{X}$ is perpendicular to $\vec{A} - \vec{B}$, or

$$(\vec{X} - \vec{C}) \cdot (\vec{A} - \vec{B}) = 0 \; .$$

From the first two relations, on addition we find

$$0 = \vec{X} \cdot (\vec{B} - \vec{C}) + (C - A) - \vec{A} \cdot \vec{B} + \vec{A} \cdot \vec{C} - \vec{B} \cdot \vec{C} + \vec{A} \cdot \vec{B} \; ,$$

or

$$\vec{X} \cdot (\vec{B} - \vec{A}) + \vec{C} \cdot (\vec{A} - \vec{B}) = 0$$

so that

$$(\vec{X} - \vec{C}) \cdot (\vec{B} - \vec{A}) = 0 \; ,$$

as we sought to prove.

Example 11-4b. Given a line ℓ and a point X not on ℓ find the point Y on ℓ such that XY is perpendicular to ℓ .

We write the equation of the line parametrically as

$$\vec{Z} = (1 - t)\vec{A} + t\vec{B} \; ,$$

so that the desired point Y is given as

$$\vec{Y} = (1 - t)\vec{A} + t\vec{B} \; .$$

The line from $\vec{X}$ to $\vec{Y}$ is perpendicular to ℓ if and only if

$$(\vec{X} - \vec{Y}) \cdot (\vec{A} - \vec{B}) = 0 \; ,$$

or

$$(\vec{X} - \vec{A}) \cdot (\vec{A} - \vec{B}) + t(\vec{A} - \vec{B}) \cdot (\vec{A} - \vec{B}) = 0 \; ,$$

so that we may solve for t , and then obtain

$$\vec{Y} = \frac{(\vec{X} - \vec{B}) \cdot (\vec{A} - \vec{B})\vec{A} + (\vec{A} - \vec{X}) \cdot (\vec{A} - \vec{B})\vec{B}}{|\vec{A} - \vec{B}|^2} \; .$$

Example 11-4c. A characterization of the plane by means of the dot product.

We describe a plane in the previous section in terms of three vectors (whose endpoints lie) in the plane. We now give an alternate description in terms of a position vector for a point in the plane and any vector $\vec{N}$ perpendicular (or normal) to the plane.

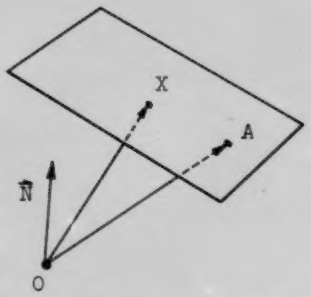


Figure 11-4d

Let A and X by any two points of the plane (Figure 11-4d), then $\vec{X} - \vec{A}$ is parallel to a line segment in the plane, hence perpendicular to $\vec{N}$; consequently,

$$\vec{N} \cdot (\vec{X} - \vec{A}) = 0 . \tag{8}$$

Conversely, if $(\vec{X} - \vec{A}) \cdot \vec{N} = 0$, then $\vec{X} - \vec{A}$ is a vector parallel to the plane and hence the point X given by

$$\vec{X} = (\vec{X} - \vec{A}) + \vec{A}$$

lies in the plane since A does. If we write (8) in terms of coördinates $\vec{X} = (x,y,z)$ and $\vec{N} = (a,b,c)$, and set $\vec{N} \cdot \vec{A} = d$, then (8) becomes

$$ax + by + cz = d ,$$

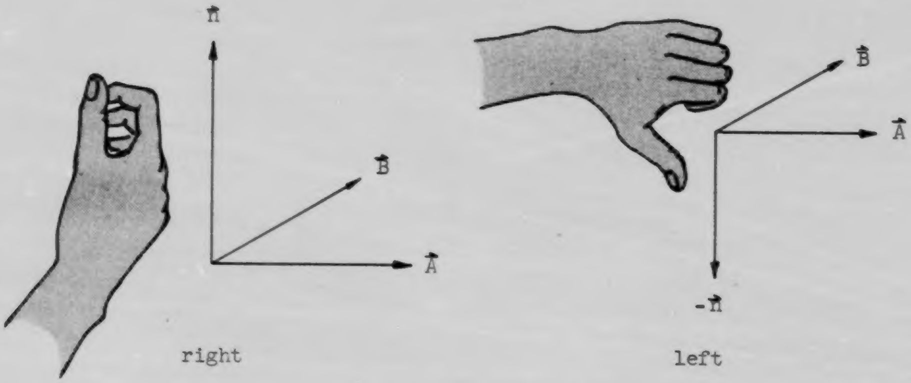
which is the familiar equation for a plane.

(ii) The cross product. We shall now define a product of two vectors which is a vector. For this product of the vectors $\vec{A}$ and $\vec{B}$ we use the parallelogram with adjacent sides $\overline{OA}$ and $\overline{OB}$. The area of the parallelogram is invariant under rotation of the coordinate frame. To associate a vector with the parallelogram we use a normal to the plane of the figure. If θ is the angle between $\vec{A}$ and $\vec{B}$, $(0 \leq \theta \leq \pi)$, $|\vec{A}||\vec{B}| \sin \theta$ is the area of the parallelogram. Let $\vec{n}$ be a unit vector perpendicular to the plane of the parallelogram. We introduce the vector

$$\vec{U} = A \otimes B = (|\vec{A}| \cdot |\vec{B}| \sin \theta)\vec{n} . \tag{9}$$

If $\vec{A}$ and $\vec{B}$ are collinear then the "parallelogram" is not defined and hence its normal vector $\vec{n}$ is not defined; but then at least one of $|\vec{A}|$, $|\vec{B}|$, $\sin \theta$ is 0 and we take $\vec{U} = \vec{0}$ in (11). In what follows we assume $\vec{A}$ and $\vec{B}$ are noncollinear unless an explicit statement to the contrary is made.

There are two unit vectors perpendicular to a given plane so that (9) does not uniquely define a vector $\vec{U}$ unless we make a specific choice for the "half-space" into which the vector $\vec{n}$ points. We choose the half-space from which the rotation through the angle θ of the ray OA into the ray OB is seen as counterclockwise (recall that $0 \leq \theta \leq \pi$) . Equivalently, we choose $\vec{n}$ so that the ordered triple $(\vec{A},\vec{B},\vec{n})$ is right-handed, where $(\vec{A},\vec{B},\vec{n})$ is a right-handed triple when the rotation through the angle θ from the direction of $\vec{A}$ into that of $\vec{B}$ is indicated by the fingers of the right hand and the thumb of the right hand points in the direction of $\vec{n}$ (Figure 11-4e). If the fingers of the left hand were used to indicate the rotation, then the left thumb would point in the direction opposite to $\vec{n}$. The two situations are depicted in the figure.



right

left

Figure 11-4e

By the foregoing rule-of-thumb if $(\vec{A},\vec{B},\vec{n})$ is right-handed, then so is $(\vec{B},\vec{A},-\vec{n})$ so that

$$\vec{A} \otimes \vec{B} = -\vec{B} \otimes \vec{A} \ . \tag{10}$$

With the convention that the null vector is parallel to every vector, note that $\vec{A}$ and $\vec{B}$ are parallel if and only if

$$\vec{A} \otimes \vec{B} = 0 \ . \tag{11}$$

Note also, since $\vec{A} \otimes \vec{B}$ is perpendicular to both $\vec{A}$ and $\vec{B}$, that

$$(\vec{A} \otimes \vec{B}) \cdot \vec{A} = (\vec{A} \otimes \vec{B}) \cdot \vec{B} = 0 \ .$$

Next we seek a coordinate representation for $\vec{A} \otimes \vec{B}$. At this point it pays to do a little wishful thinking. We have not proved the linearity properties prescribed at the beginning of this section, but let us proceed as though we had and verify afterward that our work is sound. We choose a right-handed coordinate system; that is we choose mutually perpendicular set of axes so the unit vectors $\vec{i}$, $\vec{j}$ and $\vec{k}$ in the directions of the x , y and z axes, respectively, constitute a right-handed triple in the given order (Figure 11-4f). Such a triple $(\vec{i},\vec{j},\vec{k})$ is called a fundamental set.

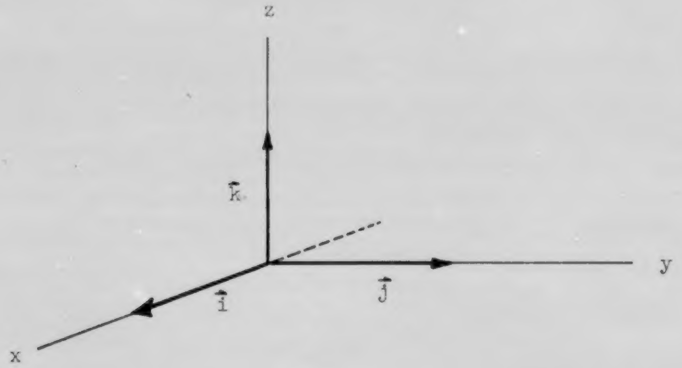


Figure 11-4f

Observe that

(12) $$\vec{i} \otimes \vec{j} = \vec{k} \, , \; \vec{j} \otimes \vec{k} = \vec{i} \, , \; \vec{k} \otimes \vec{i} = \vec{j} \, ,$$

and also

(13) $$\vec{i} \otimes \vec{i} = \vec{j} \otimes \vec{j} = \vec{k} \otimes \vec{k} = 0 \, .$$

Since $\vec{i} = (1,0,0)$, $\vec{j} = (0,1,0)$, $\vec{k} = (0,0,1)$ the coordinate representation of a vector can be written in the form

(14) $$\vec{W} = (W_x, W_y, W_z) = W_x \vec{i} + W_y \vec{j} + W_z \vec{k} \, ,$$

namely, as a linear combination of the vectors of the fundamental set. Now, we apply the assumed linearity properties to

$$\vec{A} \otimes \vec{B} = (A_x \vec{i} + A_y \vec{j} + A_z \vec{k}) \otimes (B_x \vec{i} + B_y \vec{j} + B_z \vec{k})$$

and obtain, with the aid of (12) and (13), the product

(15) $$\vec{V} = \vec{A} \times \vec{B} = (A_y B_z - A_z B_y)\vec{i} + (A_z B_x - A_x B_z)\vec{j} + (A_x B_y - A_y B_x)\vec{k} \, ,$$

where we must prove that $\vec{A} \times \vec{B}$ as defined by (15) is, in fact, the vector $\vec{A} \otimes \vec{B}$ defined by (9). It is clear that the product $\vec{V}$ defined by (15) has the stated linearity properties.

The coordinate representation (15) defines a vector $\vec{V}$ in a given coordinate frame. This vector is called the cross product of $\vec{A}$ and $\vec{B}$. We have yet to show that the coordinate representation given by Formula (15) defines the same vector in any other coordinate frame. This will be accomplished by showing that $\vec{V} = \vec{A} \otimes \vec{B} = \vec{U}$; namely that $\vec{V}$ is perpendicular to both $\vec{A}$ and $\vec{B}$, that $|\vec{V}| = |\vec{U}|$, and also that $(\vec{A},\vec{B},\vec{V})$ is right-handed triple of vectors. To prove that $\vec{A}$ and $\vec{V}$ are perpendicular we calculate the dot product:

$$\vec{A} \cdot \vec{V} = A_x(A_yB_z - A_zB_y) + A_y(A_zB_x - A_xB_z) + A_z(A_xB_y - A_yB_x) = 0 .$$

In the same way, it follows that $\vec{B}$ and $\vec{V}$ are perpendicualr. To prove $|\vec{V}| = |\vec{U}|$ observe first that

$$\begin{aligned} |\vec{U}|^2 = |\vec{A}|^2|\vec{B}|^2 \sin^2\theta &= |\vec{A}|^2|\vec{B}|^2(1 - \cos^2\theta) \\ &= |\vec{A}|^2|\vec{B}|^2 - (\vec{A}\cdot\vec{B})^2 . \end{aligned}$$

It is a straightforward exercise to write out the coordinate representation for this expression and verify that it is the same as the coordinate representation for $|\vec{V}|^2$ as obtained from (15).

We are left with the problem of showing that $(\vec{A},\vec{B},\vec{V})$ is a right-handed triple, where $\vec{V}$ is defined in a right-handed coordinate system by (15).

So far we know only that $\vec{V} = \pm\vec{U}$. To prove that the plus sign is the correct one requires some details which we take for granted here, but which can be proved easily in another context. In particular, we assume that any right-handed fundamental set can by a continuous rotation be transformed into any other right-handed fundamental set. Let t be the parameter by which this continuous motion is described. We have already shown that $\vec{V} = \pm\vec{U}$, where the sign depends on the coordinate system, hence on t. Thus we may set $\vec{V} = f(t)\vec{U}$ where f may have only the values $+1$ or -1. (We are assuming $\vec{U} \neq 0$.) Now the components of $\vec{U}$ and $\vec{V}$ both depend continuously* on t and, therefore, $f(t)$ must be continuous. We conclude that only one of the values $+1$ or -1 is possible (Exercises 11-4, No. 9). This means that the triple $(\vec{A},\vec{B},\vec{V})$ must be always right-handed or always left-handed and we must determine which. For this purpose, we choose a special right-handed fundamental set $\{\vec{i}_0,\vec{j}_0,\vec{k}_0\}$: we choose $\vec{i}_0$ in the direction of $\vec{A}$, $\vec{j}_0$ in the plane of $\vec{A}$ and $\vec{B}$ so that the sense of rotation in the plane from $\vec{A}$ to $\vec{j}_0$ is the same as that from $\vec{A}$ to $\vec{B}$, and $\vec{k}_0$ perpendicular to the plane of $\vec{A}$ and $\vec{B}$ (Figure 11-4g). In this special coordinate system, $\vec{A} = (A_x,0,0)$ where

---

*The precise concept of continuous vector function is defined in the next section. For the present, we need assume only that the components of the vector function are continuous.

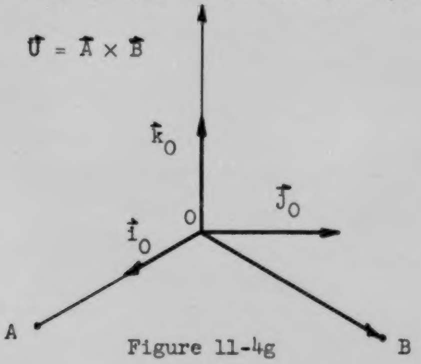


Figure 11-4g

$A_x = |\vec{A}| > 0$, and $\vec{B} = (B_x,B_y,0)$ where $B_y > 0$, (see Exercises 11-4, No. 12). Now, by (15), $\vec{V} = (0,0,A_xB_y)$; hence $V_z > 0$. The vector $\vec{V}$ has the same direction as $\vec{k}_0$ and we conclude, for this special frame, hence for all right-handed coordinate systems, that the vector $\vec{V}$ defined by (15) is $\vec{U} = \vec{A} \otimes \vec{B}$.

Vectors were introduced for the purpose of describing properties independent of the coordinate system. To better appreciate this idea and to understand the reason why so much effort has gone into the description of the vector product, we make a change of coordinates in which a right-handed fundamental set $\{\vec{i},\vec{j},\vec{k}\}$ is replaced by the left-handed fundamental set $\{\vec{i},\vec{j},\vec{k}^*\}$ where $\vec{k}^* = -\vec{k}$. If the coordinate representation of a vector $\vec{A}$, in the right-handed system is (A_x,A_y,A_z), then in the left-handed system it is $(A_x,A_y,-A_z)$. If we use Formula (15) in the right-handed and left-handed systems to describe $\vec{V}$ and $\vec{V}^*$, respectively, we obtain the following comparison, where U_x^*, U_y^*, U_z^* are the components of $\vec{U}$ in the left-handed system

right-handed system	left-handed system
$V_x = U_x$	$V_x^* = -U_x = -U_x^*$
$V_y = U_y$	$V_y^* = -U_y = -U_y^*$
$V_z = U_z$	$V_z^* = U_z = -U_z^*$

Thus, the entity defined by Formula (15) does not transform like a vector. It is a pseudo-vector; that is, it transforms like a vector under rotations of coordinate axes, but when the orientation of a coordinate axis is reversed the components transform not into components of the corresponding vector, but into their negatives. However, it is more convenient to have the same formula for the cross product in all coordinate systems than to change the formula when the orientation of an axis is reversed. For our purposes we need consider only rotations and we do not reverse coordinate axes. We restrict ourselves to right-handed coordinate systems and make no distinction between vectors and pseudo-vectors. There are physically meaningful quantities of both types, however, and in some contexts it is necessary to distinquish the two.

Example 11-4d. The normal to a plane. We have two characterizations of a plane, first, by any noncollinear triple A , B , C of its points (Formula (4) of Section 11-3) and, second, by two vectors, one normal to the plane, the other a position vector for any point in the plane (Example 11-4c). We now obtain a normal to the plane in the form

$$\vec{N} = (\vec{B} - \vec{A}) \times (\vec{C} - \vec{A})$$

and thus find a normal in terms of the noncollinear triple of points.

Example 11-4e. The reduction of the triple cross product $\vec{A} \times (\vec{B} \times \vec{C})$.

Since $\vec{B} \times \vec{C}$ is normal to the plane of $\vec{B}$ and $\vec{C}$, and $\vec{A} \times (\vec{B} \times \vec{C})$ is a vector perpendicular to $\vec{B} \times \vec{C}$, it follows that $\vec{A} \times (\vec{B} \times \vec{C})$ lies in the plane of $\vec{B}$ and $\vec{C}$, or by the corollary in Example 11-3c,

$$\vec{A} \times (\vec{B} \times \vec{C}) = s\vec{B} + t\vec{C} .$$

Now $\vec{A}$ is perpendicular to $\vec{A} \times (\vec{B} \times \vec{C})$, or $\vec{A} \cdot (\vec{A} \times (\vec{B} \times \vec{C})) = 0$, so that

$$s\vec{A} \cdot \vec{B} + t\vec{A} \cdot \vec{C} = 0 .$$

Hence we may set

$$s = (\vec{A} \cdot \vec{C})k$$

$$t = -(\vec{A} \cdot \vec{B})k$$

or

$$\vec{A} \times (\vec{B} \times \vec{C}) = k((\vec{A} \cdot \vec{C})\vec{B} - (\vec{A} \cdot \vec{B})\vec{C}) .$$

It suffices to choose a particular coordinate system and to examine one component in this relation to verify that $k \equiv 1$, hence that

$$\vec{A} \times (\vec{B} \times \vec{C}) = (\vec{A} \cdot \vec{C})\vec{B} - (\vec{A} \cdot \vec{B})\vec{C} . \tag{16}$$

We leave it to you to verify that (16) holds for the degenerate cases ignored in the above derivation (Exercises 11-4, No. 13).

Example 11-4f. Three noncoplanar vectors $\vec{A}$, $\vec{B}$, $\vec{C}$ determine a parallelepiped with initial edges $\overline{OA}$, $\overline{OB}$, $\overline{OC}$, (Figure 11-4h). The volume of the parallelepiped is the area of the base times the altitude. The (signed) volume of the parallelepiped is the so-called "triple scalar product."

$$V = \vec{A} \cdot (\vec{B} \times \vec{C}) , \tag{17}$$

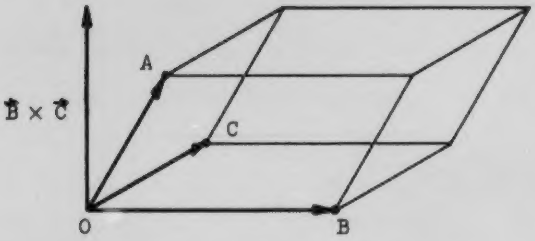


Figure 11-4h

where V is positive if $(\vec{A},\vec{B},\vec{C})$ is right-handed set, negative if left-handed (Exercises 11-4, No. 14). Thus the condition that the three vectors are coplanar is $\vec{A}\cdot(\vec{B}\times\vec{C}) = 0$.

Exercises 11-4

1. Verify properties (5a - d) of the dot product.

2. Obtain the coordinate representation (3) of the dot product.

3. (a) Show that the perpendicular projection of the vector $\vec{B}$ on the line of $\vec{A}$ is the vector

$$\vec{B}_A = \frac{|\vec{B}| \cos \theta}{|\vec{A}|} \vec{A} = \frac{\vec{B} \cdot \vec{A}}{\vec{A}^2} \vec{A} .$$

The vector $\vec{B}_A$ is called the component of $\vec{B}$ in the direction of $\vec{A}$.

(b) Write $\vec{B}$ in the form $\vec{B} = \vec{B}_A + \vec{B}^A$ and show that $\vec{B}^A$ is perpendicular to $\vec{B}_A$. The vector $\vec{B}^A$ is called the component of $\vec{B}$ perpendicular to $\vec{A}$.

4. Let $\vec{A}$ and $\vec{B}$ be noncollinear vectors. Minimize $|\vec{B} - \lambda\vec{A}|$ and interpret geometrically.

5. Prove that the diagonals of a rhombus intersect at right angles.

6. Prove that an angle inscribed in a semicircle is a right angle.

7. (a) Show that the sum of the squares of the sides of a parallelogram is equal to the sum of the squares of the diagonals.

(b) Show for an arbitrary quadrilateral that the sum of the squares of the sides exceeds the sum of the squares of the diagonals by four times the square of the distance between the midpoints of the diagonals.

8. Given $\vec{U} = (1,1,1)$, find vectors $\vec{V}$, $\vec{W}$ so that $(\vec{U},\vec{V},\vec{W})$ is a right-handed triple of mutually perpendicular vectors.

9. In connection with the definition of cross product, why must a function, continuous on an interval, which can take on only the values ± 1 be constant?

10. Show that $\vec{A} \times \vec{B} = \vec{A} \times \vec{B}^A$, (see No. 3).

11. (a) Express the vector $\vec{V} = (\vec{A}^C)^B - (\vec{A}^B)^C$ in terms of dot and cross product.

(b) Compare $\vec{V}$ with $\vec{U} = (\vec{A}_C)_B - (\vec{A}_B)_C$ and $\vec{W} = (\vec{A}_B)^C - (\vec{A}^C)_B$.

12. In the text, given noncollinear vectors $\vec{A}$ and $\vec{B}$, we chose a right-handed fundamental set $\{\vec{i}_0, \vec{j}_0, \vec{k}_0\}$ with $\vec{i}_0$ in the direction of $\vec{A}$, with $\vec{j}_0$ in the plane of $\vec{A}$ and $\vec{B}$ so that the rotation in the plane from $\vec{A}$ to $\vec{j}_0$ is in the same sense as that from $\vec{A}$ to $\vec{B}$, and with $\vec{k}_0$ perpendicular to the plane of $\vec{A}$ and $\vec{B}$. Express $\vec{i}_0$, $\vec{j}_0$, $\vec{k}_0$ in terms of $\vec{A}$ and $\vec{B}$.

13. Verify (16) for the degenerate cases ignored in its derivation.

14. Prove the result of Example 11-4f.

15. Show
$$\begin{aligned}\vec{A}\cdot(\vec{B}\times\vec{C}) &= (\vec{A}\times\vec{B})\cdot\vec{C}\\ &= -\vec{A}\cdot(\vec{C}\times\vec{B})\\ &= \vec{B}\cdot(\vec{C}\times\vec{A})\\ &= -\vec{B}\cdot(\vec{A}\times\vec{C})\\ &= \vec{C}\cdot(\vec{A}\times\vec{B})\\ &= -\vec{C}\cdot(\vec{B}\times\vec{A})\end{aligned}$$
Give a general rule for the sign?

16. Let $\vec{A}+\vec{B}+\vec{C}=\vec{0}$. Show that
$$\vec{A}\times\vec{B}=\vec{B}\times\vec{C}=\vec{C}\times\vec{A}.$$
Interpret this result geometrically to obtain the law of sines for triangles.

17. Use (16) and Number 15 to express
$$(\vec{A}\times\vec{B})\cdot(\vec{C}\times\vec{D})$$
in terms of dot products alone.

18. What is the shortest distance between the straight lines
$$\mathcal{L}_1 : \vec{X}=\vec{C}+s\vec{A}$$
$$\mathcal{L}_2 : \vec{Y}=\vec{D}+t\vec{B}\ ?$$
What is the equation of the line perpendicular to both?

19. Use cross products to find the equation of the ray which bisects the angle between $\overrightarrow{OA}$ and $\overrightarrow{OB}$. (Compare Exercises 11-3, No. 5.)

20. Prove that
$$\vec{A}\times(\vec{B}\times\vec{C})+\vec{B}\times(\vec{C}\times\vec{A})+\vec{C}\times(\vec{A}\times\vec{B})=\vec{0}.$$

21. Use (16) to find two different representations of $(\vec{A} \times \vec{B}) \times (\vec{C} \times \vec{D})$ and so establish an identity of the form $a\vec{A} + b\vec{B} + c\vec{C} + d\vec{D} = 0$. Hence, show how to express any vector as a linear combination of any three vectors $\vec{A}$, $\vec{B}$, $\vec{C}$ for which $\vec{A} \cdot (\vec{B} \times \vec{C}) \neq \vec{0}$. (Compare Exercises 11-3, No. 8).

22. Solve $\vec{X} = \vec{A} + (\vec{B} \times \vec{X})$.

11-5. Vector Calculus and Curves.

We have introduced an algebra of vectors; now we shall develop a differential calculus of vectors analogous to the calculus of ordinary functions. The vector calculus will be applied to the study of curves and, later, to mechanics.

(i) Limits, derivatives, and integrals of vector functions. We have worked with real functions on a real domain; these will now be called scalar functions to distinguish them from the functions which are our present concern, vector functions on a real domain. A vector function will be indicated by a singly barbed arrow, just as a vector is distinguished from a scalar, e.g., $\vec{r} : t \longrightarrow \vec{X}$. A vector function may be described by a representation in some fixed coordinate system

$$\vec{r}(t) = (f(t)\,, g(t)\,, h(t))$$

where f , g , h are scalar functions on a common domain. Given an origin, we may represent a vector as a point, and a vector function $\vec{r}$ as a function which assigns a point $\vec{r}(t)$ in space to each value of t in some domain. In mechanics we take t as time and the equation $\vec{X} = \vec{r}(t)$ represents the path of a particle whose position is given as a function of the time parameter. In Section 11-3, Equation (1), a line is described by means of a vector function of the parameter λ . In general, if the domain of $\vec{r}$ is an interval and $\vec{r}$ is continuous we think of the points $\vec{r}(t)$ as describing a curve.

Limit is the basic idea of the calculus. For the vector calculus, too, we shall need the idea of limit for a vector function. Since we already have, in the length of a vector, a concept analogous to that of absolute value, we may immediately extend the idea in Chapter 3, without change of form, to vectors.

DEFINITION 11-5. Let t_0 be a point for which every deleted neighborhood, contains points of the domain $\vec{r}$. We say $\vec{r}$ has the limit $\vec{A}$ at t_0 and write

(1) $$\vec{A} = \lim_{t \sim t_0} \vec{r}(t)$$

if and only if for each positive number ϵ there exists a positive number δ such that

(2) $$|\vec{r}(t) - \vec{A}| < \epsilon$$

for every t in the domain of $\vec{r}$ which satisfies

$$0 < |t - t_0| < \delta \,.$$

In (2) the ϵ-neighborhood of A, namely, $\{R : |\vec{R} - \vec{A}| < \epsilon\}$ is now a sphere about the point $\vec{A}$ of radius ϵ (Figure 11-5a). With this slight generalization of the idea of neighborhood we may describe the idea of limit geometrically, precisely as before (see Definition 3-2 and the related footnote, p. 59).

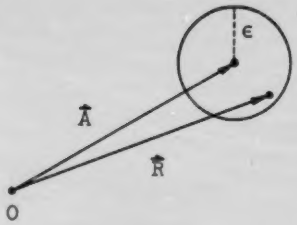


Figure 11-5a

Since the difference between two vectors does not depend on the coordinate system, it is clear that this definition of limit does not depend on the choice of coordinates.

For vectors, it is not necessary to recapitulate the epsilonic definition of limit; observe that (1) is completely equivalent to

$$\lim_{t \to t_0} |\vec{r}(t) - \vec{A}| = 0 \tag{3}$$

(Exercises 11-5, No. 1).

Some of the proofs of limit theorems in Section 3-4 can be directly transferred to prove theorems about vectors, in particular, the corollary to Theorem 3-4c for a linear combination,

$$\lim_{t \to t_0} \sum_{i=1}^{n} a_i \vec{r}_i(t) = \sum_{i=1}^{n} a_i \lim_{t \to t_0} \vec{r}_i(t) , \tag{4}$$

and Theorem 3-4d for a product,

$$\lim_{t \to t_0} a(t)\vec{r}(t) = \left[\lim_{t \to t_0} a(t)\right]\left[\lim_{t \to t_0} \vec{r}(t)\right] . \tag{5}$$

As a consequence of (4) and (5), for the coordinate representation $\vec{V}(t) = (f(t), g(t), h(t))$ it follows that $\lim_{t \to t_0} \vec{V}(t) = \vec{A}$ where $\vec{A} = (a,b,c)$ if and only if

$$\lim_{t \to t_0} f(t) = a , \lim_{t \to t_0} g(t) = b , \lim_{t \to t_0} h(t) = c . \tag{6}$$

Finally, for the products of vectors, we have

(7) $$\lim_{t \to t_0} [\vec{u}(t) \cdot \vec{v}(t)] = \left[\lim_{t \to t_0} \vec{u}(t)\right] \cdot \left[\lim_{t \to t_0} \vec{v}(t)\right] .$$

and

(8) $$\lim_{t \to t_0} [\vec{u}(t) \times \vec{v}(t)] = \left[\lim_{t \to t_0} \vec{u}(t)\right] \times \left[\lim_{t \to t_0} \vec{v}(t)\right] .$$

The proofs of these limit theorems are left to Exercises 11-5, Number 2.

With the definition of limit in hand, and the algebra of limits established, we may proceed to develop vector calculus.

A vector function $\vec{r}$ is said to be continuous at a point t_0 of its domain if

$$\lim_{t \to t_0} \vec{r}(t) = \vec{r}(t_0) .$$

Continuity on an interval corresponds to the geometrical idea of a curve without gaps.

The derivative $\vec{r}' = D\vec{r}$ of the vector function $\vec{r}$ is defined by

(9) $$\vec{r}'(t) = \lim_{\tau \to t} \frac{\vec{r}(\tau) - \vec{r}(t)}{\tau - t} .$$

In Chapter 2 we introduced the velocity of a straight line motion as the derivative of position with respect to time. Position in space can be described as a vector $\vec{r}(t)$; for the motion of a particle in space a full definition of the vector velocity is

$$\vec{v}(t) = \vec{r}'(t) .$$

In Chapter 12, we shall have much more to say about velocity $\vec{v}(t)$ and acceleration $\vec{a}(t) = \vec{v}'(t)$; for the moment we merely note these as possible applications of the concept of derivative.

The following elementary properties of derivatives are direct consequences of (9).

Let $\vec{r}$ be given by $\vec{r}(t) = (f(t), g(t), h(t))$ in some coordinate system; then $\vec{r}$ is differentiable if and only if f , g , and h are differentiable, and

(10) $$\vec{r}'(t) = (f'(t), g'(t), h'(t)) .$$

If f , $\vec{u}$, and $\vec{v}$ are differentiable, then

(11) $$D_t(f(t)\vec{u}(t)) = f'(t)\vec{u}(t) + f(t)\vec{u}'(t) ,$$

$$D_t(\vec{u}(t) \cdot \vec{v}(t)) = \vec{u}'(t) \cdot \vec{v}(t) + \vec{u}(t) \cdot \vec{v}'(t) \ , \tag{12}$$

$$D_t(\vec{u}(t) \times \vec{v}(t)) = \vec{u}'(t) \times \vec{v}(t) + \vec{u}(t) \times \vec{v}'(t) \ . \tag{13}$$

$$D_t\vec{u}(f(t)) = \vec{u}'(f(t))f'(t) \ . \tag{14}$$

The proofs of (10) - (14) are left to Exercises 11-5, Number 3.

Example 11-5a. If $|\vec{r}(t)|$ is constant, then $\vec{r}(t)$ and $\vec{r}'(t)$ are perpendicular.

By the hypothesis,

$$0 = D_t[|\vec{r}(t)|^2] = D_t[\vec{r}(t) \cdot \vec{r}(t)] = 2\vec{r}(t) \cdot \vec{r}'(t) \ ;$$

hence, $\vec{r}(t) \cdot \vec{r}'(t) = 0$ which proves the assertion.

Example 11-5b. If $\vec{r}'(t) = \vec{0}$ for all t, then $\vec{r}(t)$ is a constant.

This result follows at once from (10) and the analogous result for scalar functions (Corollary 1 to Theorem 5-4a).

Example 11-5c. If $\vec{r}(t) \times \vec{r}''(t) = 0$ then $\vec{r}(t) \times \vec{r}'(t) = \vec{k}$ where $\vec{k}$ is a constant vector.

This result is very useful in mechanics; it follows immediately from the result of Example 11-5b upon the observation that $\vec{r}(t) \times \vec{r}''(t) = D[\vec{r}(t) \times \vec{r}'(t)]$.

Example 11-5d. If $\vec{r}'(t) = f(t)\vec{A}$ and $\vec{r}(t_0) = \vec{B}$, then

$$\vec{r}(t) = \vec{B} + \left[\int_{t_0}^{t} f(\tau)d\tau\right]\vec{A} \ . \tag{15}$$

The vector function $\vec{r}$ defined by (15) satisfies the stated vector differential equation and initial condition. Furthermore, if $u(t)$ is any other solution of this initial value problem, then

$$\vec{u}'(t) - \vec{r}'(t) = f(t)\vec{A} - f(t)\vec{A} = 0 \ ,$$

so that $\vec{u}'(t) - \vec{r}'(t) = \vec{0}$, and, by Example 11-5b, $\vec{u} - \vec{r}$ is a constant function. Since $\vec{u}(t_0) - \vec{r}(t_0) = 0$, it follows that $\vec{u}(t) = \vec{r}(t)$ for all t, hence that the solution (15) is unique.

For the present, we shall not need a definition of the integral, $\int_a^b \vec{r}(t)dt$, although you will have no difficulty in defining it by analogy with the limit of Riemann sums (Exercises 11-5, No. 4). We merely note that in a coordinate representation the calculus of vector functions is equivalent component-by-component to the calculus of scalar functions. We have seen this in (6) and (10), and it remains true for the integrals of vectors as well.

(ii) Parametric representations of curves. As we have already mentioned, a continuous vector function $\vec{r}$ on an interval defines a curve in space. The equation $\vec{X} = \vec{r}(t)$ is called a parametric representation of the curve. In a coordinate representation with $\vec{X} = (x,y,z)$, such a parametric representation has the form

$$\begin{cases} x = f(t) \\ y = g(t) \\ z = h(t) . \end{cases} \tag{16}$$

For each value of t , the function $\vec{r}$ determines a point $\vec{X} = (x,y,z)$ and the set of points defines the curve.

For the purposes of this text we restrict ourselves to piecewise smooth curves; i.e., curves for which the function $\vec{r}$ in a parametric representation $\vec{X} = \vec{r}(t)$ is not only continuous, but also has a piecewise continuous derivative. We emphasize that a curve has not one but infinitely many parametric representations. Two parametric representations $\vec{X} = \vec{r}(t)$ and $\vec{X} = \vec{q}(u)$ of the same curve are said to be equivalent if on the domain of $\vec{q}$ there exists a scalar function ϕ , continuous, increasing, and piecewise continuously differentiable, for which the range of ϕ is the domain of r and $r(\phi(u)) = q(u)$. Clearly, given any parametric representation $X = r(t)$, it is possible to describe infinitely many others by composition with appropriate functions ϕ .

Suitable choices of parameter will often yield very simple representations of complicated curves, even curves which intersect themselves or which may have sharp spikes (cusps).

Example 11-5e. The circle.

To represent the circle $(x - a)^2 + (y - b)^2 = c^2$ by means of the graphs of continuous functions we must split the curve into two semi-circles:

$$y = b + \sqrt{c^2 - (x - a)^2} , \qquad (a - c \leq x < a + c) ,$$

and

$$y = b - \sqrt{c^2 - (x - a)^2}\ , \qquad (a - c < x \leq a + c)\ ,$$

while the parametric representation

$$x = a + c \cos \theta\ ,\ y = b + c \sin \theta$$

for $0 \leq \theta \leq 2\pi$ provides a one-to-one map of the parameter interval onto the entire circle.

In order to give a geometrical interpretation to $\vec{r}'(t)$, we consider the definition

$$\vec{r}'(t_0) = \lim_{t \to t_0} \frac{\vec{r}(t) - \vec{r}(t_0)}{t - t_0}$$

and observe (Figure 11-5b) that the vector $\vec{X} - \vec{X}_0$, where $\vec{X}_0 = \vec{r}(t_0)$ and $\vec{X} = \vec{r}(t)$, is represented by a directed chord of the curve. If the direction of this chord has a limit as t approaches t_0 it is that of the tangent line. Consequently, if $\vec{r}'(t_0) \neq \vec{0}$, the vector $\vec{T}_0 = \vec{r}'(t_0)$, attached to the point X_0 , is tangent to the curve there. From its definition, the vector $\vec{T}_0$ points along the curve in the direction of increasing t . If $\vec{r}'(t_0) = \vec{0}$ then the derivative does not define a direction for the curve at X_0 (although a direction might exist in fact, as we shall soon see).

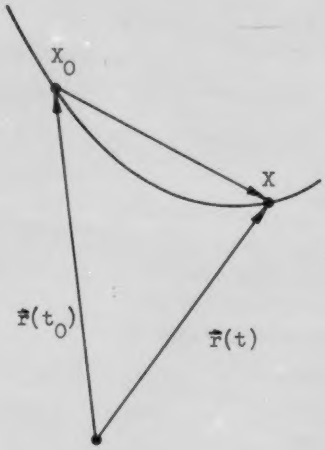


Figure 11-5b

If $\vec{r}'(t) \neq 0$, it is useful to single out the unit tangent vector

$$\vec{t} = \frac{\vec{r}'(t)}{|\vec{r}'(t)|} \tag{17}$$

since, apart from a possible reversal of direction, $\vec{t}$ does not depend on the choice of parameter (the use of t for the parameter and $\vec{t}$ for the unit tangent vector is awkward but conventional.) Hereafter, when we refer to "the tangent" without qualitifcation we shall mean the unit tangent vector (17).

Example 11-5f. The semicubical parabola.

Consider the "semicubical parabola" defined by

$$x = t^2 \ , \ y = t^3 \ .$$

We may split the curve into the graphs of two function:

$$y = x^{3/2} \ , \ y = -x^{3/2} \ , \ (x \geq 0)$$

(Figure 11-5c). In the figure, the arrows indicate the direction of increasing t along the curve. At the origin the curve has a cusp, a point where the curve has a discontinuous reversal of direction, although the parametric representation is smooth. More precisely, since $\vec{r}'(t) = (2t, 3t^2)$, the tangent is

$$t = \left(\frac{2t}{\sqrt{4t^2 + 9t^4}} , \frac{3t^2}{\sqrt{4t^2 + 9t^4}} \right) .$$

Consequently $\lim\limits_{t \sim 0^+} \vec{t} = (1,0)$, while $\lim\limits_{t \sim 0^-} \vec{t} = (-1,0)$.

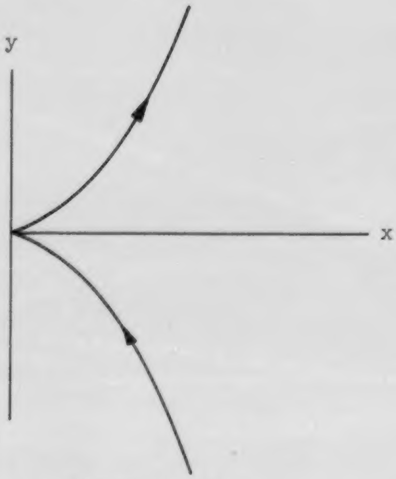


Figure 11-5c

A null derivative may be the fault only of some peculiarity of the parametrization rather than of the curve, as the following example shows.

Example 11-5g. The straight line.

We have already given the parametric representation of the straight line

$$\vec{r}(t) = \vec{A} + \vec{B}t \ ,$$

thus $\vec{r}(s) = \vec{A} + \vec{B}s^3$ is the same straight line. At the point A corresponding to $s = 0$, $\frac{d\vec{r}}{ds} = 0$. Nonetheless

$$\lim_{s \to 0} \vec{t} = \lim_{s \to 0} \frac{\vec{B}}{|\vec{B}|} = \frac{\vec{B}}{|\vec{B}|} ,$$

so that the tangent vector $\vec{t}$ is defined for all s . At A , the curve is perfectly regular; that is, there exists a parametrization of the line $\vec{X} = q(\sigma)$ for which $\vec{A} = \vec{q}(\sigma_0)$ and $\vec{q}'(\sigma_0) \neq \vec{0}$.

If $\vec{r}'(t) = \vec{0}$ at a point then the tangent vector and tangent direction may not exist. We study only curves and parametrizations for which the zeros of $\vec{r}'$ are isolated; that is, each value of t for which $\vec{r}'(t) = \vec{0}$ is contained in a neighborhood with no other zeros of $\vec{r}'$. Whenever we change parameter we require that this condition be maintained. Where $\vec{r}'(t)$ vanishes the curve may have a cusp, or there may be regular point of the curve, or there may be more complicated behavior.

The arclength of a plane curve was introduced in Section 6-3(iv). We parallel the discussion there to obtain a formula for the arc length of any curve given parametrically. Let the curve have the parametric representation $\vec{X} = \vec{r}(t)$ for $a \leq t \leq b$ where $\vec{r}'(t)$ is continuous on $[a,b]$. Let $\sigma = [t_0, t_1, \ldots, t_n]$ be any partition of $[a,b]$. The length of the polygon joining the successive points $\vec{r}(t_i)$, $(i = 0,1,\ldots,n)$ is

$$(18)\quad P(\sigma) = \sum_{k=1}^{n} |\vec{r}(t_k) - \vec{r}(t_{k-1})|$$

$$= \sum_{k=1}^{n} \sqrt{[f(t_k) - f(t_{k-1})]^2 + [g(t_k) - g(t_{k-1})]^2 + [h(t_k) - h(t_{k-1})]^2}\ ,$$

where f, g, h are the component functions in the coordinate representation of $\vec{r}$. By the Law of the Mean for each of the component functions, there exist numbers ξ_k, η_k, ζ_k in (t_{k-1}, t_k) such that

$f(t_k) - f(t_{k-1}) = f'(\xi_k)(t_k - t_{k-1})$, $g(t_k) - g(t_{k-1}) = g'(\eta_k)(t_k - t_{k-1})$, and $h(t_k) - h(t_{k-1}) = h'(\zeta_k)(t_k - t_{k-1})$. Consequently,

$$P(\sigma) = \sum_{k=1}^{n} \sqrt{f'(\xi_k)^2 + g'(\eta_k)^2 + h'(\zeta_k)^2}\,(t_k - t_{k-1})\ .$$

It follows from the continuous differentiability of $\vec{r}$ (hence of f, g, and h) that the limit of $\lim\limits_{\nu(\sigma)\to 0} P(\sigma)$ exists and is the Riemann integral

$$(19)\qquad L = \int_a^b \sqrt{f'(t)^2 + g'(t)^2 + h'(t)^2}\,dt$$

$$= \int_a^b |\vec{r}'(t)|dt\ .$$

The proof of this result is left to Exercises 11-5, Number 13.

It is easy to verify that the arclength defined by the integral formula is the same for all equivalent parametrizations of the curve (given by a substitution $t = \phi(u)$ where ϕ is increasing and ϕ' continuous). Furthermore, from the vector representation of L, the arclength is independent of the coordinate system. These properties make arclength a very convenient parameter for the general theory of curves. For this purpose we introduce the parameter s, the arclength measured along the curve from a fixed reference point $\vec{r}(t_0)$, by

$$s = \psi(t) = \int_{t_0}^{t} |r'(\tau)|d\tau \,. \tag{20}$$

Since $\vec{r}'(t)$ vanishes at isolated points, if at all, the function ψ is increasing, and the derivative ψ' is continuous and may have zeros only at isolated points. Thus s and t are equivalent parameters. Except for places where $\vec{r}'(t) = 0$ we may invert the relation and consider t as a function of s with $\frac{dt}{ds} = \frac{1}{|\vec{r}'(t)|}$. Given a representation of the curve $\vec{X} = \vec{r}(t)$, we have

$$\frac{d\vec{X}}{ds} = \frac{d\vec{X}}{dt}\frac{dt}{ds} = \frac{1}{|\vec{r}'(t)|}\frac{d\vec{X}}{dt} \,;$$

hence,

$$\left|\frac{d\vec{X}}{ds}\right| = 1 \tag{21}$$

and, by (17), $\frac{d\vec{X}}{ds}$ is the tangent to the curve at each point.

Although it is extremely convenient to parametrize curves in terms of arclength to prove general theorems, it is often impossible to invert the relation $s = \psi(t)$ for a given curve and we must rely on the original representation $\vec{r}(t)$ to obtain explicit results.

Example 11-5h. The Four Cusped Hypocycloid.

Consider the curve given by

$$\begin{cases} x = \cos^3\theta \\ y = \sin^3\theta \end{cases}, \qquad (0 \le \theta \le 2\pi) \,.$$

From (20)

$$s = \int_0^{\theta} 3|\cos\phi \sin\phi|d\phi \,,$$

$$s = \begin{cases} \frac{3}{2}\sin^2\theta & , \text{ for } \ 0 \leq \theta \leq \frac{\pi}{2} \\ 3 - \frac{3}{2}\sin^2\theta & , \text{ for } \ \frac{\pi}{2} \leq \theta \leq \pi \\ 3 + \frac{3}{2}\sin^2\theta & , \text{ for } \ \pi \leq \theta \leq \frac{3\pi}{2} \\ 6 - \frac{3}{2}\sin^2\theta & , \text{ for } \ \frac{3\pi}{2} \leq \theta \leq 2\pi \ . \end{cases}$$

In each case, $\frac{ds}{d\theta} = 3|\cos\theta\sin\theta|$, and

$$\vec{t} = (\cos\theta, \sin\theta)\ \mathrm{sgn}\ (\cos\theta\sin\theta) \ .$$

Thus the curve has cusps at $\theta = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$, (Figure 11-5d)

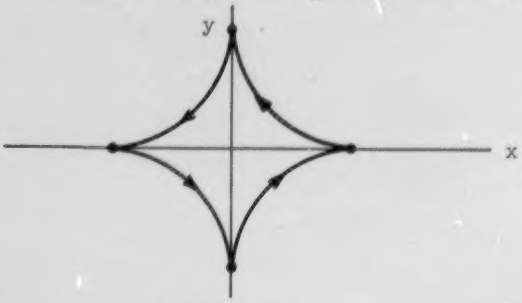


Figure 11-5d

Example 11-5i. Now we prove the geometrically reasonable proposition that if all tangent lines to a curve are concurrent then the curve is a straight line. Let the curve be given by $\vec{X} = \vec{r}(s)$. The equation of the tangent line at the particular point X is

$$\vec{Y} = \vec{X} + \lambda\frac{d\vec{X}}{ds} ,$$

where λ is the parameter on the line. If A is the common point of the tangent lines; then for each s there is a value $\lambda = \phi(s)$ such that

$$\vec{A} = \vec{r}(s) + \phi(s)\vec{r}\,'(s) \ . \tag{22}$$

Take the x-component in this equation, to obtain from $\vec{r}(s) = (f(s), g(s), h(s))$,

$$f'(s)\phi(s) + f(s) = A_x \ .$$

This is a linear differential equation of first order (Section 10-7). Since the constant function $s \longrightarrow A_x$ is a particular solution, the solution can be put in the form

$$f : s \longrightarrow A_x + B_x \exp\ \{-\int_0^A \frac{d\sigma}{\phi(\sigma)}\}$$

where B_x is the constant $f(0) - A_x$. Since the solution has the same form for each component, we obtain the solution of (22) in the form

$$\vec{X} = \vec{r}(s) = \vec{A} + \omega(s)\vec{B} \tag{22}$$

where $\vec{B} = \vec{X}_0 - \vec{A} = \vec{r}(0) - \vec{A}$. We have only to replace $\omega(s)$ by a parameter μ to see that (23) is the equation of a straight line.

In this example we see again how the calculus of scalar functions, this time exemplified by the solution of a linear differential equation, can be transformed component by component into the vector calculus.

Exercises 11-5

1. Prove that Definition 11-5 and Equation (3) are equivalent definitions of limit for a vector function.

2. (a) Prove Properties (4) and (5) for the limits of vector functions.

   (b) Use the results of Part (a) to prove Property (6).

   (c) Prove Properties (7) and (8).

3. Prove the vector differentiation formulas (10) - (14).

4. Consider the function $\phi : t \longrightarrow |\vec{r}(t)|$ where $\vec{r}$ is differentiable. Differentiate

   (a) $\phi(t)$.

   (b) $\dfrac{1}{\phi(t)}$.

   (c) $\phi(t)^2$.

5. Obtain the formula for the arclength of a plane curve given in polar coordinates by $\rho = f(\theta)$.

6. Sketch each of the following curves, give the points at which the tangent vector doesn't exist, and represent the curve in terms of arclength where possible. (If no z-coordinate is given, restrict the locus to the x,y-plane).

   (a) $\begin{cases} x = a + b \sin t \\ y = c + d \cos t \end{cases}, \qquad (0 \le t \le 2\pi)$.

   (b) The cycloid,

   $$\begin{cases} x = a(t - \sin t) \\ y = a(1 - \cos t) \end{cases}, \qquad (a > 0 \ , \ 0 \le t \le 2\pi) \ .$$

   Show that this curve is the locus of a point on a circle that rolls on a straight line without slipping.

   ∧ (c) The Cornu spiral,

   $$\begin{cases} x = \displaystyle\int_0^t \cos u^2 \, du \\ y = \displaystyle\int_0^t \sin u^2 \, du \ . \end{cases}$$

(d) The cardioid,

$$\begin{cases} x = \cos\theta\,(1 - \cos\theta) \\ y = \sin\theta\,(1 - \cos\theta) \end{cases}, \qquad (0 \le \theta \le 2\pi),$$

or, in polar coordinates, $\rho = 1 - \cos\theta$.

(e)
$$\begin{cases} x = \dfrac{2t}{1 + t^2} \\[2ex] y = \dfrac{1 - t^2}{1 + t^2} \end{cases}.$$

Identify this curve.

(f) The three-dimensional helix,

$$\begin{cases} x = a\cos t \\ y = a\sin t \\ z = t \end{cases}, \qquad (a > 0).$$

(g) The conical helix,

$$\begin{cases} x = a\,t\cos t \\ y = a\,t\sin t \\ z = t \end{cases}. \qquad (a > 0).$$

7. Show how to define the integral $\int_a^b \vec{r}(t)dt$ by the method of Riemann sums. Prove that this is equivalent to integrating component by component.

8. What is the unique continuously differentiable solution of $\vec{r}'(t) = t\vec{A} + \vec{B}$ with the initial condition

$$\vec{r}(0) = \vec{0}\ ?$$

9. (a) Let the parametric representation of a curve be given in the form $\vec{X} = \vec{r}(s)$ where s is arclength and $\vec{r}$ is three times differentiable. From $|\vec{t}| = \left|\frac{d\vec{X}}{ds}\right| = 1$, it follows from Example 11-5a that $\frac{d\vec{t}}{ds}$ is perpendicular to $\vec{t}$. If $\frac{d\vec{t}}{ds} \ne 0$, the unit vector $\vec{n} = \frac{d\vec{t}}{ds} \Big/ \left|\frac{d\vec{t}}{ds}\right|$ exists. The vector $\vec{n}$ is called the principle normal to the curve. Assuming that $\vec{n}$ exists for all s , prove that the curve is planar if and only if

$$\vec{t} \times \frac{d\vec{n}}{ds} = 0 .$$

(b) Express this condition in terms of derivatives of $\vec{X}$.

10. We have restricted ourselves to parametrizations of curves $\vec{X} = \vec{r}(t)$, $a \leq t \leq b$, for which the derivative $\vec{r}'$ is continuous with isolated zeros, if any. In general, we say that $\vec{t}$ is the tangent to the curve at X_0 , if there is a parametrization $\vec{r}$, with $\vec{X}_0 = \vec{r}(t_0)$, such that

$$\vec{t} = \lim_{t \to t_0^+} \vec{v}(t) = \lim_{t \to t_0^-} \vec{v}(t) ,$$

where

$$\vec{v}(t) = \frac{\vec{r}(t) - \vec{r}(t_0)}{t - t_0} \bigg/ \left| \frac{\vec{r}(t) - \vec{r}(t_0)}{t - t_0} \right| .$$

(a) Show that this definition includes the text case

$$\vec{t} = \frac{\vec{r}'(t)}{|\vec{r}'(t)|} .$$

(b) In the text we have defined parametrizations $\vec{X} = \vec{q}(\tau)$ and $\vec{X} = \vec{r}(t)$ as equivalent if $\vec{r}(\phi(\tau)) = \vec{q}(\tau)$ where ϕ is defined on the domain of $\vec{q}$, has a range in the domain of $\vec{r}$, is increasing, and has a piecewise continuous derivative. Prove that the tangent $\vec{t}$ at X_0 as defined in Part (a) is the same for all equivalent parametrizations.

(c) If, in the definition of equivalent parameters ϕ is replaced by a decreasing function, we say that t and τ are "contravalent" parameters, just to have a word for it. Show that contravalent parametrizations orient the curve in opposite senses; that is, if $\vec{t}$ is the tangent for the parametrization $\vec{X} = \vec{r}(t)$, then $-\vec{t}$ is the tangent for $\vec{X} = \vec{q}(\tau) = \vec{r}(\phi(\tau))$.

11. A possible vector generalization of Rolle's Theorem is: Let $\vec{r}$ be differentiable on $a \leq t \leq b$ and let $\vec{r}(a) = \vec{r}(b) = 0$, then there is a point t , $a < t < b$, at which $\vec{r}'(t) = \vec{0}$. Prove or disprove.

Λ 12. For $\vec{X} = \vec{r}(t)$ where $\vec{r}$ has a continuous derivative on $[a,b]$, prove that the lengths $P(\sigma)$ of inscribed polygons given by (18) have a least upper bound and that this upper bound is the arclength L given by (19).

Λ13. Complete the proof that the arclength integral (19) is the limit of the lengths of inscribed polygons (18) by establishing the following lemma.

Let $\vec{R}(t) = (F(t)\ ,\ G(t)\ ,\ H(t))$ be continuous on $[a,b]$. For each partition $\sigma = \{t_0, t_1, \ldots, t_n\}$ of $[a,b]$ and each choice of ξ_k , η_k , ζ_k in $[t_{k-1}, t_k]$, $(k = 1 , \ldots , n)$, consider

$$P = \sum_{k=1}^{n} \sqrt{F(\xi_k)^2 + G(\eta_k)^2 + H(\zeta_k)^2}\ (t_k - t_{k-1}) \ .$$

Under the stated condition,

$$\lim_{\nu(\sigma)\to 0} P = \int_a^b \sqrt{F(t)^2 + G(t)^2 + H(t)^2}\ dt \ .$$

11-6. Curves in the Plane.

In this section, we make use of the vector calculus to represent geometrical quantities in an invariant way; that is, independently of changes in the coordinates. First, we give an invariant description of the area of a region in terms of a vector representation of its boundary curve. Next, we study the curvature of a plane curve at a point as a measure of rate of bending or turning. We shall see that a plane curve is completely described by giving its curvature as a function of arclength; in this way we shall have obtained a description of a curve which is not only independent of the coordinate system, but also independent of the parametrization.* We shall also study certain special kinds of curves which are not only interesting for their own sake, but have value for applications.

(i) The area enclosed by a curve. The description of area we have utliized so far is certainly not independent of the coordinate system. It was based on the idea of "standard region", the region under a curve $y = f(x)$, above the x-axis, and bounded by the two vertical lines $x = a$ and $x = b$. If we wish to obtain an invariant description of the area enclosed by a curve, we may proceed in either of two ways: either we introduce a new definition of area not related to coordinates and then show the new and old definitions agree, or we take our old definition and modify and extend it until we see easily that area does not depend on coordinate representations. We shoose the latter alternative, since it involves less work. We restrict ourselves to curves that do not intersect themselves, although we give several exercises intended to show what happens when a curve does intersect itself.

The idea of "self-intersection" is easily described in terms of a parametrization $\vec{X} = \vec{r}(t)$, $a \leq t \leq b$. We say the curve intersects itself (or, is not simple) if there exist two values of the parameter t_2 and t_1, at least one of them in the open interval (a,b), such that $\vec{r}(t_2) = \vec{r}(t_1)$. If the curve does not intersect itself, we say it is simple. If a curve is simple and $\vec{r}(a) \neq \vec{r}(b)$, it is called an arc. Note that the definition of simplicity permits $\vec{r}(a) = \vec{r}(b)$. We say a curve, simple or not, is closed if $\vec{r}(a) = \vec{r}(b)$. A simple closed curve, or (Jordan curve) is, therefore, a curve for which $\vec{r}(a) = \vec{r}(b)$ and for which $0 < |t_2 - t_1| < b - a$ implies $\vec{r}(t_2) \neq \vec{r}(t_1)$.

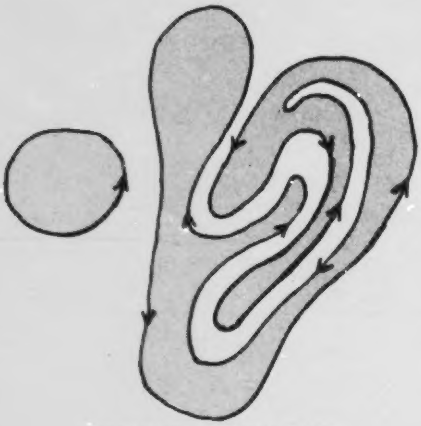
---

*Except for a reversal of orientation.

57

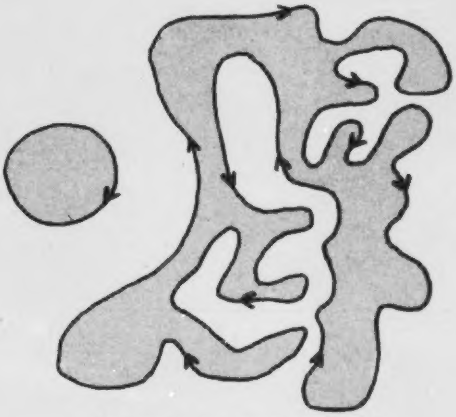
> The French mathematician, C. Jordan (1838-1922), was the first to realize that this idea of "simplicity" required deeper study. In his Cours d' Analyse (1887), Jordan stated the geometrically "obvious" theorem that a simple closed curve separates the plane into two parts, an "inside" and an "outside." Jordan considered a more general class of "curves" than we do; namely, $\vec{f}$ only need be continuous, not piecewise continuously differentiable as we assume. He attempted a proof which was rather complicated and turned out to be incomplete. This "obvious" theorem was finally proved by the American mathematician Veblen in 1905. Relatively simple proofs are now known, but they all require an extended development of geometry (topology) beyond what is feasible here.

For the moment, we confine our attention to the area of a region bounded by a simple closed curve. The parametric representation $\vec{X} = \vec{f}(t)$ defines an orientation of the curve, the direction in which the curve is traced as the parameter t is increased. With respect to this orientation it is meaningful to speak of the left and right side of the curve just as we speak of the left and right sides of a road with respect to the direction in which the road is traveled. It is useful to attach a sign to the area of the region enclosed by the curve, positive if the region lies on the left of the curve (Figure 11-6a), negative if it lies on the right (Figure 11-6b) with respect to the given orientation. At the same time, if the region enclosed by a simple closed curve lies on the left with respect to the orientation of the curve we say the orientation is positive (counterclockwise), if the region is on the right, negative (clockwise). The expression we shall obtain for the area enclosed by a plane curve is defined even if the curve intersects itself, but then we must extend our idea of area to make the result geometrically meaningful. By giving area a sign we prepare for this extension.



positive orientation

Figure 11-6a



negative orientation

Figure 11-6b

Next, we modify the definition of area for a standard region to conform with the given conventions of sign and orientation. Consider the standard region under the graph of a nonnegative function f over an interval $[a,b]$. We introduce a simple parametrization, $\vec{F}(t) = (\phi(t),\psi(t))$, $\alpha \leq t \leq \beta$, for the boundary of the standard region for which the orientation is positive. Note that the graph of f is then oriented from right to left, the reverse of the customary orientation from left to right (Figure 11-6c). Now suppose $a = \phi(t_1)$ and $b = \phi(t_0)$ where $\alpha \leq t_0 < t_1 \leq \beta$ (see Exercises 11-6, Nos. 2 and 4). Then

$$A = \int_a^b f(x)dx = \int_{t_1}^{t_0} f(\phi(t))\phi'(t)dt ;$$

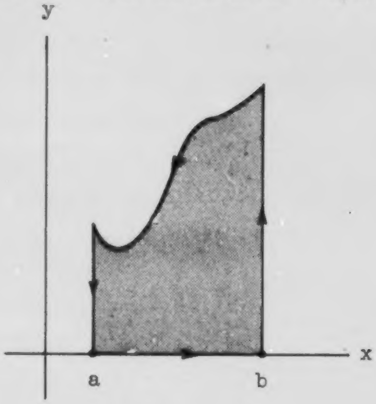


Figure 11-6c

hence

$$(1) \qquad A = -\int_{t_0}^{t_1} \psi(t)\phi'(t)dt ,$$

where the integral in (1) is taken in the direction of increasing t. Next we observe that on the straight segments either $y = 0$ or x is constant, hence $\psi(t) = 0$ or $\phi'(t) = 0$. It follows that the integral in (1) can be taken over the entire parameter interval; that is,

$$(2) \qquad A = -\int_\alpha^\beta \psi(t)\phi'(t)dt .$$

It is conventional to use Leibnizian notation and write (2) in the form

$$(3) \qquad A = -\oint yx' \, dt$$

where $y = \psi(t)$, $x = \phi(t)$ and where the orientation of the circle indicates the orientation of the simple curve; in this notation, the ends of integration are omitted.

Note that the integral (2) is defined for any allowable parametrization whether the curve is closed or simple or not. For us, the particular interest in (2) is as the area of a closed curve.

Our objective is to show that (3) agrees with our conception of area for the region enclosed by a simple closed curve for a class of curves which will include almost all cases of practical interest. Having done that, we shall simply accept (3) as a definition of area whenever the integral exists. In particular (3) is defined for the class of curves which have the piecewise continuously differentiable parametrizations to which we restrict ourselves (these curves are called piecewise smooth, smooth meaning continuously differentiable). The form (3) does not reveal independence of the coordinate system, but we shall obtain a related expression for area which does.

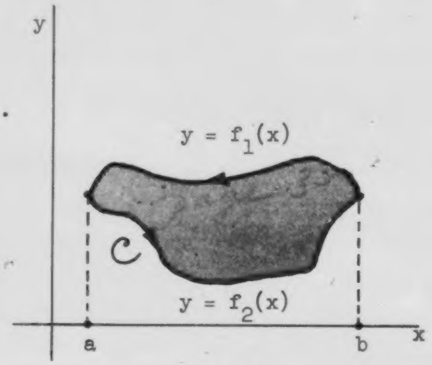


Figure 11-6d

First, as in the Introduction (Section 1-2), we consider a simple closed curve $\mathcal{C}$ which can be subdivided into finitely many arcs, each the graph of a function. The simplest case is a subdivision into two such arcs (Figure 11-6d). Let $y = f_1(x)$ represent the upper arc, and $y = f_2(x)$ the lower, where $[a,b]$ is the common domain of f_1 and f_2. Consider a positive parametrization $x = \phi(t)$, $y = \psi(t)$ for $\alpha \leq t \leq \beta$ where $\phi(\alpha) = \phi(\beta) = a$. The decomposition of $\mathcal{C}$ into two such arcs amounts to a division of the parameter interval into two subintervals, first $[\alpha\ \gamma]$ on which ϕ is increasing, then $[\gamma,\beta]$ on which ϕ is decreasing; here $\phi(\gamma) = b$. Now consider the area enclosed by $\mathcal{C}$. This is the difference between the areas of the graphs of the standard regions under the graphs of f_1 and f_2; namely,

$$A = \int_a^b f_1(x)dx - \int_a^b f_2(x)dx .$$

From (1), we have

$$A = -\int_\gamma^\beta \psi(t)\phi'(t)dt - \int_\alpha^\gamma \psi(t)\phi'(t)dt$$

where we take account of the negative orientation of the standard region under the graph of f_2 fixed by the orientation of $\mathcal{C}$ (see Exercises 11-6, No. 3). Consequently,

$$A = -\oint y\ x'dt$$

in agreement with (3).

More generally, consider a simple closed curve $\mathcal{C}$ in a positive parametrization $x = \phi(t)$, $y = \psi(t)$, $\alpha \leq t \leq \beta$, where $[\alpha,\beta]$ can be subdivided by a partition $\{t_0,t_1,\ldots,t_n\}$ into intervals on which ϕ is either strongly monotone or constant. The curve $\mathcal{C}$ can be subdivided, accordingly, into arcs which represent functions, and vertical segments (Figure 11-6e). Let $X_k = (\phi(t_k),\psi(t_k))$, $(k = 0,1,\ldots,n)$, be the points of $\mathcal{C}$ corresponding to the partition and draw the vertical line through each point X_k (the same line may pass through several of the X_k). Let A be any point above the curve which is not on one of these lines. Let us start at A and go vertically downward. When we first meet an arc of the curve we pass from the exterior to the interior, so the arc is oriented from right to left. In crossing the next arc we pass from the interior to the exterior, so the curve is traversed from left to right. In this way, we proceed across pairs of oppositely oriented arcs until we finally cross the bottom arc and pass from the interior to the exterior of $\mathcal{C}$ for the last time. Consider the region contained between each such pair of arcs and the vertical lines on either side of A, for example, the shaded region between the arcs BC and DE in Figure 11-6e. The area of this subregion is easily seen to be

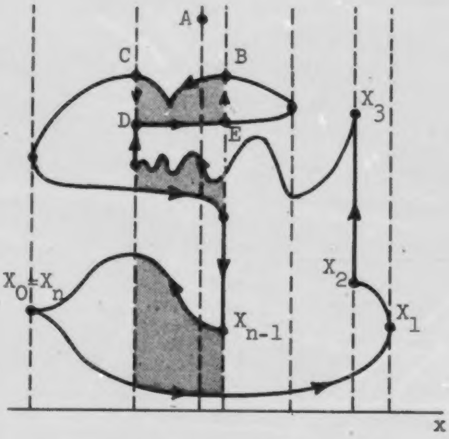


Figure 11-6e

$$-\oint_{BCDE} yx'dt = -\int_{BC} yx'dt$$

$$-\int_{DE} yx'dt$$

since the integrals over the vertical segments CD and DE vanish by the argument for (3). To find the total area cut out of the interior of $\mathcal{C}$ by the strip containing the point A we only have to add up the integrals $-\int yx'dt$ over all the arcs of $\mathcal{C}$ within the strip. Summing over all strips we see that the area enclosed by $\mathcal{C}$ is the sum of the integrals taken over all arcs where ϕ is strongly monotone. However, the integral over a vertical segment is zero, so that we may write

$$A = -\sum_{k=1}^{n}\int_{t_{k-1}}^{t_k} yx'dt = -\oint_C yx'dt \ .$$

Thus Formula (3) is established for a very general class of curves.

We have seen that (3) conforms with the ideas of area developed in Chapter 6. At the same time the integral in (3) certainly exists for a larger class than those we have just discussed, namely, for the piecewise smooth curves which we are taking as the basic class. We simply adopt (3), as a definition of area for the interior of a simple closed curve under the stated conventions of orientation and sign. It is not very hard to prove with due allowance for these conventions, that this definition satisfies the properties of area of Section 6-1, provided all regions considered have piecewise smooth curves as boundaries. Here, we simply accept that it can be done and turn our attention to the question of invariance.

Formula (3) is not symmetric in x and y . We shall need a more symmetrical form, and obtain it by getting the formula for area referred to the y-axis and taking the average of the two. We have

$$A = -\oint yx'dt = -\int_{\alpha}^{\beta} yx'dt$$

where $[\alpha,\beta]$ is the parameter interval. Integrating by parts to invert the roles of x and y (compare the geometrical discussion to Section 10-4), we obtain

$$A = -xy\Big|_{\alpha}^{\beta} + \int_{\alpha}^{\beta} xy'dt \ ,$$

but, since $\vec{r}(\alpha) = \vec{r}(\beta)$, the first term is zero and we have

$$A = \oint xy'dt \ . \tag{4}$$

Now we take the average of the expressions for A given by (3) and (4) to obtain

$$A = \frac{1}{2}\oint (xy' - yx')dt \ , \tag{5}$$

the desired form. We shall use this form to prove the invariance of area under changes in the coordinate system, but is also frequently the convenient form for the direct computation of area.

Example 11-6a. The ellipse

$$x = a + b \cos t$$
$$y = c + d \sin t , \qquad (0 \le 1 < 2\pi) ,$$

has the area

$$A = \frac{1}{2} \int_0^{2\pi} \{(a + b \cos t) d \cos t \, dt + (c + d \sin t) b \sin t\} dt$$
$$= \frac{1}{2} \int_0^{2\pi} (a\, d \cos t + c\, b \sin t) dt + \frac{bd}{2} \int_0^{2\pi} (\cos^2 t + \sin^2 t) dt$$
$$A = \pi\, bd .$$

Thus the area is π times the product of the semimajor and semiminor axes.

Example 11-6b. If a curve is given in polar form by the equation $\rho = f(\theta)$, then θ may be taken as the parameter:

$$\vec{X} = \vec{r}(\theta) = (f(\theta) \cos \theta , f(\theta) \sin \theta) .$$

From (5), we obtain the formula

$$(6) \qquad A = \frac{1}{2} \oint f(\theta)^2 \, d\theta = \frac{1}{2} \oint \rho^2 \, d\theta = \frac{1}{2} \oint \vec{X}^2 d\theta .$$

To see that the area as defined by (5) is independent of the coordinate system, imbed the plane in space. Let $\vec{i}$ and $\vec{j}$ be the unit coordinate vectors in the x, y-plane and set $\vec{v} = \vec{i} \times \vec{j}$. The vector $\vec{v}$ is the unit upward normal to the plane, (Figure 11-6f). Equation (5) can be written in the vector form

$$(7) \qquad A = \oint \vec{v} \cdot (\vec{X} \times \vec{X}') dt$$

where $\vec{X}' = \vec{r}'(t)$. Since $\vec{v}$ is unaffected by translations or rotations of the coordinate frame this is clearly the desired invariant expression for the area.

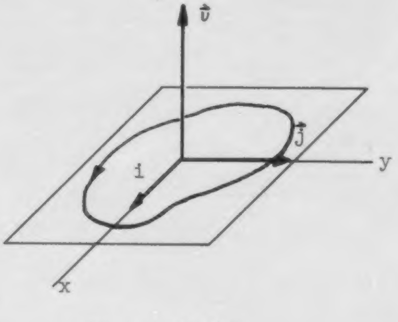


Figure 11-6f

(ii) Curvature. As a first approximation to a curve in the neighborhood of a point, we have already made use of the tangent line (Section 5-7). If we want further information in the form of a measure of the deviation of the curve from straightness we may try to formulate an analytical answer to the geometrical question, "How much is the curve bending?" As a curve bends the tangent turns and we may use the rate of turning of the tangent per unit arclength, which may be thought of as the rate at which the direction of the curve changes, to measure the bending of the curve. This implies that we must restrict ourselves to points $\vec{X} = \vec{f}(s)$ where the second derivative with respect to arclength

$$\frac{d^2\vec{X}}{ds^2} = \frac{d\vec{t}}{ds}$$

exists. If θ is the angle of inclination of $\vec{t}$ with respect to the x-axis, then $\vec{t} = (\cos\theta, \sin\theta)$, defines θ uniquely. The instanteneous rate of change of direction,

$$\kappa = \frac{d\theta}{ds} \tag{8}$$

is called the curvature at the given point. Since $\vec{t}$ has a fixed (unit) length, the vector $\frac{d\vec{t}}{ds}$ is perpendicular to $\vec{t}$ (Example 11-5a). We introduce the unit normal to the curve directed to the left of the tangent

$$\begin{aligned}\vec{n} &= (\cos[\theta + \tfrac{\pi}{2}], \sin[\theta + \tfrac{\pi}{2}]) \\ &= (-\sin\theta, \cos\theta) .\end{aligned} \tag{9}$$

We have, at once

$$\frac{d\vec{t}}{ds} = \kappa\vec{n} \tag{10}$$

and

$$\frac{d\vec{n}}{ds} = -\kappa\vec{t} . \tag{11}$$

Although we defined the curvature κ in terms of θ, we ended with an interpretation of κ, Equation (10), in terms of the tangent and left pointing normal. Thus (10) gives an invariant definition of κ, independent of the coordinate system, and we take (10) as basic rather than (8).

Note that the opposite orientation of the curve reverses the direction of both $\vec{t}$ and $\vec{n}$ in (10) so that κ does not depend upon the orientation of the curve.

We shall show later that if κ is a continuous function of arclength s then it characterizes the curve completely: namely, for any continuous function f there exists only one curve, apart from rotations and translations, such that $\kappa = f(s)$.

Example 11-6c. For the circle

$$\begin{cases} x = a + R \cos \theta \\ y = b + R \sin \theta \,, \end{cases}$$

we have

$$\frac{d\vec{r}}{d\theta} = R(-\sin \theta \,, \cos \theta) \;,$$

so that arclength is given by $s = R\theta$. With arclength as parameter, we have

$$\begin{cases} x = a + R \cos \frac{s}{R} \\ y = b + R \sin \frac{s}{R} \end{cases}$$

$$\frac{d\vec{r}}{ds} = (-\sin \frac{s}{R}, \cos \frac{s}{R}) = \vec{t}(s)$$

$$\frac{d\vec{t}}{ds} = -\frac{1}{R}(\cos \frac{s}{R}, \sin \frac{s}{R}) = \frac{\vec{n}(s)}{R}$$

where $\vec{n}(s) = (-\cos \frac{s}{R}, -\sin \frac{s}{R})$. Thus the curvature of a circle is constant, consistent with our intuition that the tangent to a circle turns at a constant rate. The radius of the circle is the reciprocal of the curvature.

For an arbitrary curve we define the radius of curvature R as the reciprocal of the curvature. The significance of the radius of curvature will be further explored below.

Example 11-6d. The straight line.

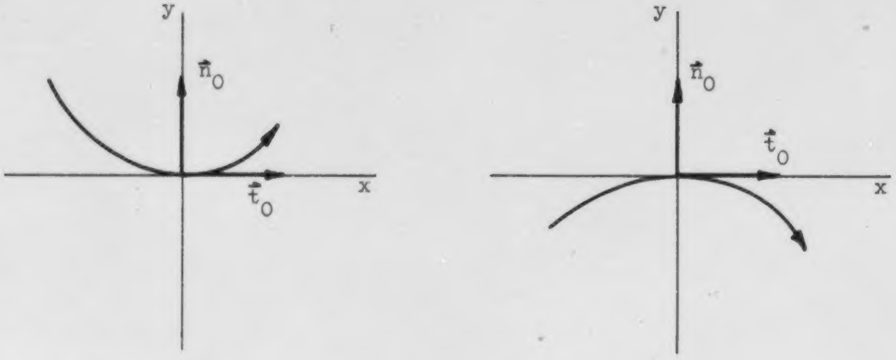
The equation for a straight line in terms of arclength is

$$\vec{r}(s) = \vec{a} + s \frac{\vec{b}}{|\vec{b}|} \;.$$

Consequently, $\frac{d\vec{t}}{ds} = 0$ and hence $\kappa = 0$. Conversely, if $\kappa = 0$ then $\vec{t}$ is constant and $\vec{r}(s)$ describes a straight line.

Example 11-6e. The relation between sign of curvature and flexure of a curve.

Let $\vec{X}_0 = \vec{r}(s_0)$ be any point of the curve and choose a coordinate frame with origin at X_0 , x-axis in the direction of $\vec{t}_0$, and y-axis in the direction of $\vec{n}_0$ (Figure 11-6g).



positive curvature negative curvature

Figure 11-6g

Now let us take the y-components in the equations $\vec{t} = \frac{d\vec{X}}{ds}$ and $\frac{d\vec{t}}{ds} = \frac{d^2\vec{X}}{ds^2} = \kappa\vec{n}$ at $s = s_0$, to obtain

$$\left.\frac{dy}{ds}\right|_{s_0} = 0$$

and

$$\left.\frac{d^2y}{ds^2}\right|_{s_0} = \kappa_0 .$$

If $\kappa_0 > 0$ it follows that y_0 is a local minimum, if $\kappa_0 < 0$ that y_0 is a local maximum. Thus, in the neighborhood of X_0, the curve lies on the same side of the tangent as the normal, or on the opposite side, according to whether κ is positive or negative. In the language of Section 5-5, we say that the curve is flexed at X_0 in the direction of $\kappa\vec{n}$.

Example 11-6f. Representation of curvature in terms of a parameter other than arclength.

Let the curve be $\vec{X} = \vec{f}(t)$, where t need not be arclength. Using primes to indicate differentiation with respect to t, we have

$$s' = |\vec{X}'| \quad \text{and} \quad \vec{t} = \vec{X}' \Big/ |\vec{X}'| \ ;$$

whence

$$\frac{d\vec{t}}{ds} = \vec{t}' \Big/ |\vec{X}'| = \frac{1}{|\vec{X}'|} \frac{d}{dt}\left(\frac{\vec{X}'}{|\vec{X}'|}\right)$$

$$= \frac{\vec{X}'}{|\vec{X}'|^2} - \frac{(\vec{X}'' \cdot \vec{X}')\vec{X}'}{|\vec{X}'|^4},$$

where, at the last step, we have used

$$\frac{d}{dt}\frac{1}{|\vec{X}'|} = -\frac{1}{|\vec{X}'|^2}\frac{d}{dt}\sqrt{\vec{X}' \cdot \vec{X}'} = -\frac{\vec{X}'' \cdot \vec{X}'}{|\bar{X}'|^3} .$$

(compare Exercises 11-5, No. 4(b)). We recognize $\frac{d\vec{t}}{ds}$ as the component of $\frac{\vec{X}''}{|\vec{X}'|^2}$ normal to $\vec{X}'$. If θ is the angle between $\vec{X}''$ and $\vec{X}'$, then $|\kappa| = \left|\frac{d\vec{t}}{ds}\right| = \frac{|\vec{X}''|}{|\vec{X}'|^2}\sin\theta = \frac{|\vec{X}' \times \vec{X}''|}{|\vec{X}'|^3}$ where the sign of κ is positive if $\vec{X}''$ points to the left of the tangent and negative if $\vec{X}''$ points to the right. Consequently, if we introduce the unit upward normal to the plane $\vec{N} = \vec{t} \times \vec{n}$, we have

$$\kappa = \frac{\vec{N} \cdot (\vec{X}' \times \vec{X}'')}{|\vec{X}'|^3} . \tag{12a}$$

From this we easily obtain the coordinate representation of κ: for $\vec{X} = (x,y)$,

$$\kappa = \frac{x'y'' - y'x''}{[(x')^2 + (y')^2]^{3/2}} . \tag{12b}$$

In particular if the curve is the graph of a function $f : x \longrightarrow y$ we may take x as the parameter and obtain

$$\kappa = \frac{f''(x)}{[1 + f'(x)^2]^{3/2}} \tag{12c}$$

thus κ is positive or negative according to whether the curve is flexed upward or downward (compare Exercises 5-8, No. 11(a)). Since we can always choose the tangent line as the x-axis, it follows from (12c) that the curvature is zero and changes sign at an inflection point.

(iii) Center of curvature. Evolute and involute. To improve the approximation to the graph of a function in the neighborhood of a given point by a line through the point which has the same direction as the curve, the tangent line, we may take account of the flexure of the curve, also, and approximate it by a circle through the point which has the same tangent and curvature, the so-called osculating circle.* This would be the circle with center situated at distance equal to the radius of curvature along the normal. The center $\vec{Y}$ of the osculating circle is called the center of curvature, and, by definition, is given by

$$\vec{Y} = \vec{X} + \frac{1}{\kappa}\,\hat{n} = \vec{X} + R\hat{n} \tag{13}$$

where X is the point of osculation on the curve and $R = \frac{1}{\kappa}$ is the radius of curvature.

There is another way of defining the osculating circle and center of curvature which is geometrically appealing. Let the curve be given by $\vec{X} = \vec{r}(s)$ and let $\vec{X}_0 = \vec{r}(s_0)$ where the osculating circle is sought. Consider any three points on the curve $\vec{X}_1 = \vec{r}(s_1)$, $\vec{X}_2 = \vec{r}(s_2)$, and $\vec{X}_3 = \vec{r}(s_3)$ where s_1 , s_2 , and s_3 are distinct. Consider the circle through the three points. As s_1 , s_2 , and s_3 all approach s_0 , this circle approximates a limiting circle, the osculating circle we have defined above. To prove that this is so requires techniques which will be better understood after we have studied Taylor's Theorem in Chapter 13. Just as the tangent line gives the "best" linear approximation (compare Section 5-7) to the curve near X_0 (in the sense that the error of approximation of any other line is greater than that of the tangent line for all sufficiently small neighborhoods of s_0) the osculating circle gives the "best" approximation to the curve of all circles.

Equation (13) defines a new curve, the locus of the centers of curvature, $\vec{Y} = \vec{q}(s)$ where $\vec{q} : s \longrightarrow \vec{X} + R\hat{n}$. This locus is called the evolute of the original curve, and the original curve is called an involute of the new locus. The parameter s in $\vec{Y} = \vec{q}(s)$ is, of course, not the arclength for the evolute, but a convenient parameter. The evolute and involutes of a given curve are useful in many physical applications, and we shall meet some of them in Chapter 15.

There is a second geometrical approach to the idea of evolute which is particularly illuminating. For this we consider the center of curvature in a somewhat different light, as the limit of the intersections of normal lines. For a circle, all normal lines intersect at the center. We anticipate, then, that as s approaches s_0 , the intersection of the normal lines through X and X_0 approaches the center of the osculating circle. To prove this result, write the equation of the two normals:

---

[*] From Latin, osculor, to kiss.

$$\vec{U} = \vec{X}_0 + \lambda\vec{n}_0$$
$$\vec{V} = \vec{X} + \mu\vec{n} \ .$$

If the normal lines are not parallel $(\vec{n} \neq \vec{n}_0)$ the lines intersect at a point $\vec{U} = \vec{V}$ such that

$$\vec{X} - \vec{X}_0 = \lambda\vec{n}_0 - \mu\vec{n} \ .$$

We may eliminate μ and solve for λ by taking the cross-product with $\vec{n}$:

$$\lambda(\vec{n} \times \vec{n}_0) = \vec{n} \times (\vec{X} - \vec{X}_0) \ .$$

Observe that this equation is equivalent to

$$\lambda\left(\frac{\vec{n} - \vec{n}_0}{s - s_0} \times \vec{n}_0\right) = \vec{n} \times \left(\frac{\vec{X} - \vec{X}_0}{s - s_0}\right) \ .$$

We obtain the limit of λ , hence the limit of the point of intersection, as s approaches s_0 , by

$$\lambda \left.\frac{d\vec{n}}{ds}\right|_{s_0} \times \vec{n}_0 = \vec{n}_0 \times \left.\frac{d\vec{X}}{ds}\right|_{s_0}$$

or, by (11),

$$-\lambda\kappa(t_0 \times n_0) = n_0 \times t_0 \ ;$$

whence $\lambda = \frac{1}{\kappa} = R$, as we sought to prove.

The construction we have just described is a special case of the concept, envelope of a family of curves. Let a family of curves be given such that each value of a parameter α defines a curve $\vec{U} = \vec{r}(t,\alpha)$. The envelope $\vec{Y} = \vec{q}(\alpha)$ of the family is the locus of the limit of intersections of the curves $\vec{U} = \vec{r}(t,\alpha)$ and $\vec{V} = \vec{r}(t,\beta)$ as β approaches α . The preceding discussion demonstrates that the evolute is the envelope of the normal lines. The reason for the name "envelope" or "outer wrapper" is easily appreciated from Figure 11-6h which shows a curve and its evolute. We shall consider the envelope of a general family of straight lines in a later example.

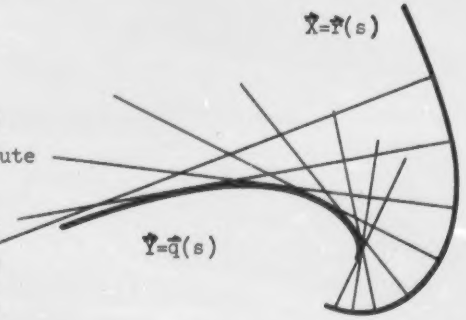


Figure 11-6h

In order to examine the relation between a curve $\vec{X} = \vec{r}(s)$ and its evolute $\vec{Y} = \vec{q}(s)$, we express the evolute in the form $\vec{Y} = \vec{\rho}(\sigma)$ where σ is arclength for the evolute, and we denote the tangent and normal on the evolute by $\vec{\tau}$ and $\vec{\nu}$, respectively. Now, we differentiate with respect to s in (13) to obtain with the aid of (11)

$$\frac{d\vec{Y}}{ds} = -\frac{1}{\kappa^2}\frac{d\kappa}{ds}\vec{n} \ . \tag{14}$$

Consequently, the normal to the curve is the tangent to the evolute when $\frac{d\kappa}{ds}$ is negative, and is the negative of that tangent when $\frac{d\kappa}{ds}$ is positive; thus,

$$\vec{\tau} = -\left(\operatorname{sgn}\frac{d\kappa}{ds}\right)\vec{n} \ . \tag{15}$$

Since $\sigma = \left|\frac{d\vec{Y}}{ds}\right|$, we have from (14)

$$\frac{d\sigma}{ds} = \left|\frac{d}{ds}\frac{1}{\kappa}\right| = \left(\frac{d}{ds}\frac{1}{\kappa}\right)\left(\operatorname{sgn}\frac{d}{ds}\frac{1}{\kappa}\right) = -\left(\frac{d}{ds}\frac{1}{\kappa}\right)\operatorname{sgn}\frac{d\kappa}{ds} \ .$$

hence, if $\frac{d\kappa}{ds}$ has constant sign in an interval, we obtain

$$\sigma = -\frac{1}{\kappa}\operatorname{sgn}\frac{d\kappa}{ds} + c \ , \tag{16}$$

where c is constant on the interval.

The formula for σ exhibits two kinds of singularity. When $\kappa = 0$, "infinite" radius of curvature, the corresponding point on the evolute is undefined. Thus a straight line has no evolute. When $\frac{d\kappa}{ds} = 0$, the evolute is again singular. For example the evolute of a circle is its center, a point, not a curve. More typically, the condition $\frac{d\kappa}{ds} = 0$ corresponds to an extremum of the curvature. We shall see below that this corresponds to a cusp for the evolute.

The inverse problem, given a curve, to find an involute, also has its uses. Let $\vec{Y} = \vec{\rho}(\sigma)$ be the given curve and assume $\vec{X} = \vec{r}(s)$ is an involute. We suppose that the correspondence between s and σ given by Equation (16) is one-to-one. Then, from (13) and (15),

$$\vec{X} = \vec{Y} + \frac{1}{\kappa}\left(\operatorname{sgn}\frac{d\kappa}{ds}\right)\frac{d\vec{Y}}{d\sigma} \ ;$$

hence, from (16),

$$\vec{X} = \vec{Y} + (c - \sigma)\frac{d\vec{Y}}{d\sigma} \ ; \tag{17a}$$

where c is constant. Consequently, if $\vec{Y} = \vec{\rho}(\sigma)$ has an involute, the involute can be parametrized by

$$\vec{X} = \vec{\rho}(\sigma) + (c - \sigma)\vec{\rho}\,'(\sigma) \; . \tag{17b}$$

On the other hand, for each constant c, (17b) defines a curve unambiguously. It is a simple exercise to verify that $\vec{Y} = \vec{\rho}(\sigma)$ is, in fact, the evolute for each such curve (Exercises 11-6, No. 15). We see then, that a given curve has infinitely many different involutes corresponding to different choices of the constant c; i.e., different initial points for the measurement of σ.

Example 11-6g. Envelopes of straight lines and the Clairaut differential equation.

Consider the family of straight lines

$$y = mx + f(m) \tag{18}$$

where each member of the family is defined by its slope m. Since $\frac{dy}{dx} = m$ the members of this family satisfy the differential equation

$$y = x\,\frac{dy}{dx} + f\left(\frac{dy}{dx}\right) , \tag{19}$$

which is known as the Clairaut Equation. Now we may ask whether the family of straight lines comprises all the solutions of (19). If we could find a new curve which has the members of the family (18) as its tangents, then at each point of the curve it would have the same slope as the line in the family which passes through that point; then the new curve also must satisfy the Clairaut Equation. We shall show that the envelope of the family of straight lines (provided one exists) is just such a curve.

Let $y = m_0x + f(m_0)$ be a fixed line of the family and let $y = mx + f(m)$ be any other line. The point of intersection of the two lines is given by $m_0x + f(m_0) = mx + f(m)$ or $x = -\dfrac{f(m) - f(m_0)}{m - m_0}$. Consequently the limit point of intersection as m approaches m_0 is given by $x = -f'(m)$. Thus, a point (x,y) on the envelope of the family satisfies the equations

$$\begin{cases} x = -f'(m) \\ y = -mf'(m) + f(m) , \end{cases}$$

which may be taken as the parametric representation of the envelope. In any interval where the function f' is strongly monotone we may use x as the parameter. Thus, if μ is the inverse of $-f'$, then $m = \mu(x)$ and $y = x\mu(x) + f(\mu(x))$. On differentiation with respect to x, we obtain

$$\begin{aligned}\frac{dy}{dx} &= \mu(x) + x\mu'(x) + f'(\mu(x))\mu'(x)\\ &= \mu(x) + x\mu'(x) - x\mu'(x)\\ &= \mu(x)\ .\end{aligned}$$

Replacing $\mu(x)$ by $\frac{dy}{dx}$ in the expression for y we see that the envelope satisfies the Clairaut Equation (19).

Example 11-6h. Cusps of the evolute and involute.

Let $\vec{X} = \vec{r}(s)$ be the given curve, and, in the notation we have been employing, let $\vec{Y} = \vec{\rho}(\sigma)$ be its evolute. It can be proved that an isolated extremum of the curvature for the given curve corresponds to a cusp of the evolute. We prove the result under a slightly more restrictive condition. Consider a point $\vec{X}_0 = \vec{r}(s_0)$ where $\kappa \neq 0$ (the center of curvature is defined), $\frac{d\kappa}{ds} = 0$ (the curvature is stationary at s_0), and $\frac{d^2\kappa}{ds^2} \neq 0$ (the curvature at s_0 is an extremum). In (15) we then observe that

$$\operatorname{sgn}\frac{d\kappa}{ds} = \operatorname{sgn}\left[(s - s_0)\frac{d^2\kappa}{ds^2}\right]$$

for any s in some neighborhood of s_0 (proof of Theorem 5-5a); hence $\lim_{s\to s_0^+}\vec{\tau} = -\lim_{s\to s_0^-}\vec{\tau}$, so that the tangent to the evolute reverses direction at s_0.

The involute of $\vec{Y} = \vec{\rho}(\sigma)$ given by (17b) has the tangent $\vec{t}$ given by

$$\begin{aligned}\vec{t} &= \frac{d\vec{X}}{ds} = \frac{d\sigma}{ds}\frac{d\vec{X}}{d\sigma} = \frac{d\sigma}{ds}(c - \sigma)\frac{d\vec{\tau}}{d\sigma}\\ &= \frac{d\sigma}{ds}(c - \sigma)\gamma\,\vec{\nu}\ ,\end{aligned}$$

where γ is the curvature of $\vec{Y} = \vec{\rho}(\sigma)$. Assuming $\frac{d\sigma}{ds} \neq 0$, we see that the cusps of the involute occur when just one of factors $c - \sigma$ or γ changes sign; i.e., if $\sigma = c$ or $\vec{Y}$ is an inflection point, but not if both conditions hold simultaneously.

Example 11-6i. The involute of a circle.

In Example 11-6c we obtained the parametric representation of the circle $\vec{X} = (R\cos\frac{s}{R}, R\sin\frac{s}{R})$. If we choose $c = 0$ in Equation (17b) we obtain the parametric representation of the involute,

$$\begin{cases} x = R\cos\frac{s}{R} + s\sin\frac{s}{R} \\ y = R\sin\frac{s}{R} - s\cos\frac{s}{R} \, . \end{cases}$$

The involute is a spiral with two arms which meet in a cusp at $s = 0$.

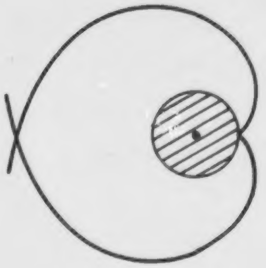


Figure 11-6i

(iv) Curvature as the characterization of a plane curve. For a given curve $\vec{X} = \vec{r}(s)$, where s is arclength, there is a unique curvature function $k : s \longrightarrow \kappa$. Conversely, given any continuous function k , there exists exactly one geometrical curve with curvature given as a function of arclength by $\kappa = k(s)$; this is the result we now prove.

If there exists a curve $\vec{X} = \vec{r}(t)$ with the curvature function κ then $\vec{r}$ must satisfy the differential equations

(20a) $$\vec{r}\,'(s) = \vec{t}$$

(20b) $$\frac{d\vec{t}}{ds} = \kappa\vec{n}$$

(20c) $$\frac{d\vec{n}}{ds} = -\kappa\vec{t}$$

and the algebraic conditions

(20d) $$|\vec{t}| = 1$$

and

(20e) $$\vec{t} \times \vec{n} = \vec{N}$$

where $\vec{N}$ is the upwardly directed normal to the plane.

A condition such as (20d) is essential for uniqueness; for without it, if $\vec{r}(t)$ were a solution of (20a,b,c) then $\lambda\vec{r}(t)$ would also be a solution where λ is any constant scalar. Condition (20d) makes it explicit that the parameter s is arclength. The condition that $\vec{N}$ is upwardly directed is also a uniqueness condition, it fixes $\vec{n}$ as the left-pointing normal. We could also fix $\vec{n}$ as the right-pointing normal, but if we did so, we would characterize not the same curve as for the left-pointing normal, but its mirror image (Figure 11-6j). We shall verify this observation in the course of the proof.

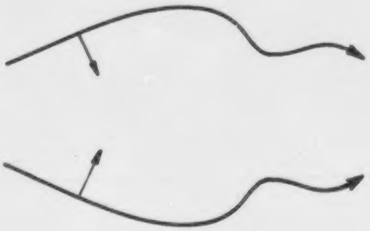


Figure 11-6j

Given $\vec{t}$ in (20a), the function $\vec{r}$ is determined directly by integration. Thus the problem is reduced to that of finding the solutions of (20b) and (20c). First, we express the conditions (20d) by writing $\vec{t}$ in the form

$$\vec{t} = (\cos\theta\,,\sin\theta),$$

where θ is a function of s . Then (20b) takes the form

(21) $$\vec{n} = \frac{1}{\kappa}\frac{d\theta}{ds}(-\sin\theta\,,\cos\theta)$$

(where we must assume $\kappa \neq 0$). Consequently, from (20c),

$$\begin{aligned}\frac{d\vec{n}}{ds} &= [\frac{d}{ds}(\frac{1}{\kappa}\frac{d\theta}{ds})](-\sin\theta\,,\cos\theta) - \frac{1}{\kappa}(\frac{d\theta}{ds})^2(\cos\theta\,,\sin\theta)\\ &= -\kappa(\cos\theta\,,\sin\theta)\,.\end{aligned}$$

Take the dot product with $\vec{t}$ to obtain $(\frac{d\theta}{ds})^2 = \kappa^2$; whence,

(22) $$\frac{d\theta}{ds} = \pm\kappa\,.$$

It follows from (21) that $\vec{n}$ is a unit normal vector. Note that there is nothing in Equations (20a-d) which fixes the sign in (22). However we now have $\bar{n} = \pm(-\sin\theta\,,\cos\theta)$ and to make $\vec{n}$ a left-pointing normal we must choose the plus sign. Integrating, we obtain

(23) $$\theta = \phi + \alpha\text{ , where }\phi = \int_0^s k(\sigma)d\sigma\text{ ,}$$

and α is any constant. We then have, at once,

$$(24) \qquad \begin{aligned} \vec{t} &= (\cos(\phi + \alpha), \sin(\phi + \alpha)) \\ \vec{n} &= (-\sin(\phi + \alpha), \cos(\phi + \alpha)) . \end{aligned}$$

Although we have assumed $k(s) \neq 0$ for all s, (24) is a solution of Equations (20b,c) for any continuous function $k : s \longrightarrow \kappa$. As often occurs in the solution of differential equations, the formal result is valid under conditions more general than those which were imposed to find it.

From the expression for $\vec{t}$ in (24) we obtain the solution of (20a) by straightforward integration. Thus, for $u : s \longrightarrow \phi + \alpha = \theta$, we have

$$(25) \qquad \begin{cases} x = x_0 + \displaystyle\int_0^s \cos u(\sigma)\, d\sigma \\ y = y_0 + \displaystyle\int_0^s \sin u(\sigma)\, d\sigma . \end{cases}$$

It is a simple exercise to verify that $\vec{r}(s) = (x,y)$ is a solution of the Equations (20a-c) which satisfies the conditions (20d,e). Observe also, that if we had replaced θ by its negative (to correspond with a right-pointing normal) in (23), the solution would have been the reflection of $\vec{X} = \vec{r}(s)$ in the line $y = y_0$, in agreement with our earlier assertion.

There are three different parameters α, x_0, y_0 in the solution (25). In what sense, then, can such a solution be unique? A change in the parameter α amounts to a rotation of the curve, and changes in x_0 and y_0 to a translation (see Exercises 11-6, No. 17(a)). Consequently, different choices of parameters in (25) merely yield the same geometrical curve in different locations in the plane. Conversely, given a parametric representation of a curve which satisfies (20), all representations obtained by rotations and translations applied to the given one are solutions of (20), (Exercises 11-6, No. 17(b)).

The question remains, are all solutions of (20) members of the family (25)? To answer this question, we fix the parameters α, x_0, y_0 by imposing conditions which specify a point on the curve and the tangent at that point initially; e.g.,

$$(26) \qquad \vec{r}(0) = \vec{X}_0 , \; \vec{r}\,'(0) = \vec{t}_0 .$$

We now prove that if the vector functions $\vec{r}$ and $\vec{\rho}$ satisfy (20) together with the initial conditions (26), that $\vec{r} = \vec{\rho}$. The implication is that there is exactly one geometrical curve defined by the curvature function k.

Let the tangents for $\vec{r}$ and $\vec{\rho}$ be denoted by $\vec{t}$ and $\vec{\tau}$, the normals by $\vec{n}$ and $\vec{\nu}$, respectively. Since the initial tangents are equal, $\vec{t}_0 = \vec{\tau}_0$, it follows that also $\vec{n}_0 = \vec{\nu}_0$. Now consider the difference vectors,

$$\vec{U} = \vec{r}(s) - \vec{\rho}(s)$$

$$\vec{V} = \vec{t} - \vec{\tau}$$

$$\vec{W} = \vec{n} - \vec{\nu} ;$$

these vectors satisfy the differential equations

(27a) $$\frac{d\vec{U}}{ds} = \vec{V}$$

(27b) $$\frac{d\vec{V}}{ds} = \kappa\vec{W}$$

(27c) $$\frac{d\vec{W}}{ds} = -\kappa\vec{V}$$

together with the initial conditions

(27d) $$\vec{U}_0 = \vec{V}_0 = \vec{W}_0 = \vec{0} .$$

By a slight variant of the argument used to prove the uniqueness of sin and cos as defined by a differential equation (Theorem 8-5b), we show that $\vec{V} = \vec{W} = \vec{0}$ for all s. Take the dot product with $\vec{V}$ in (27b), with $\vec{W}$ in (27c), and add to obtain

$$\vec{V} \cdot \frac{d\vec{V}}{ds} + \vec{W} \cdot \frac{d\vec{W}}{ds} = \frac{1}{2} \frac{d}{ds} [\vec{V}^2 + \vec{W}^2] = 0 .$$

It follows that $\vec{V}^2 + \vec{W}^2$ is constant, and from the second and third initial conditions in (27d) that the constant is $\vec{0}$. Consequently, $\vec{V} = \vec{W} = \vec{0}$. From (27a) we conclude next that $\vec{U}$ is constant and from the first initial condition, that $\vec{U} = \vec{0}$; hence $\vec{r}(s) = \vec{\rho}(s)$ and the curves coincide at every point.

Exercises 11-6

1. Verify that the following curves are simple.

   (a) The graph of a continuous function.

   (b) A circle.

   (c) A cardioid (Exercises 11-5, No. 6(d)).

   Λ(d) The Cornu spiral (Exercises 11-5, No. 6(c)).

2. Obtain a parametric representation for the boundary of the standard region under the graph of f oriented in the positive sense indicated in Figure 11-6c.

3. Show how the expression for the signed area of a standard region (1), taken in the direction of increasing t, changes if the orientation is negative.

4. In the derivation of (1) it was supposed that the part of the boundary which is the graph of the given function corresponds to a subinterval of the domain of parametrization. Show for any arc of a simple closed curve how to modify the parametrization so that the arc corresponds to a subinterval of the domain.

5. Find the area

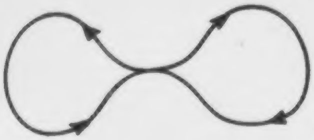
   (a) under one arch of the cycloid (Exercises 11-5, No. 6(b)).

   (b) of the interior of the cardiod (Exercises 11-5, No. 6(d)).

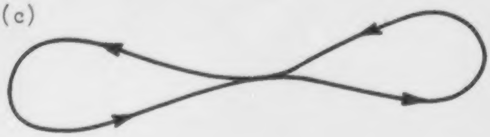
6. Sketch the curve given in polar coordinates by $\rho = a \cos \theta - b$, $a > b > 0$, for $0 \leq \theta \leq 2\pi$. Use (5) or an equivalent formula to compute the "area." The result is not the area in the usual sense. Check the derivation of (5) to see what the formula actually gives.

7. For the following curves what is the "area" as computed by (5)?

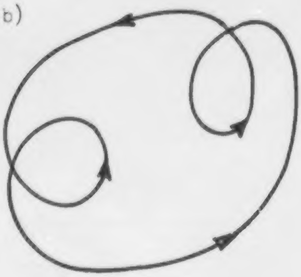
(a)



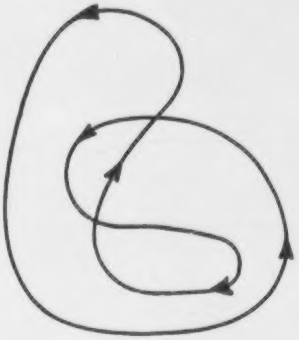
(c)



(b)



(d)



Generalize as far as you can.

8. Obtain the expression for the curvature for a curve given in polar coordinates by $\rho = f(\theta)$.

9. Find the curvature at each point of the following curves given in cartesian, polar or parametric representation as the notation suggests

(a) $y = x^2$.

(b) $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

(c) $\rho^2 = a^2 \cos 2\theta$.

(d) the cycloid (Exercises 11-5, No. 6(b)).

(e) the Cornu spiral (Exercises 11-5, No. 6(c)).

(f) the cardioid (Exercises 11-5, No. 6(d)).

(g) $\begin{cases} x = a \cos^4 \theta \\ y = a \sin^4 \theta \end{cases}$.

10. Show that the evolute of a cycloid (Exercises 11-5, No. 6(d)) is a cycloid.

11. Obtain a cartesian representation for the evolute of an ellipse and sketch the curve.

12. The involutes of a curve $\mathcal{C}$ given by $\vec{P} = \vec{\rho}(\sigma)$ can be drawn by the following simple mechanical construction. Imagine an inextensible thread wrapped tightly around the curve $\mathcal{C}$ on the side opposite the center of curvature. Cut the thread at $\sigma = c$ (compare Equation (17)) and unwrap the thread from the curve while keeping it taut. The condition of tautness requires that the unwrapped thread is pulled out straight and remains tangent to $\mathcal{C}$. Under these conditions show that the two ends of the thread at the cut describe the involute of $\mathcal{C}$.

13. The curve $\rho = ce^{a\theta}$, in polar form, has the property that the position vector of a point on the curve makes an angle with the tangent to the curve at the point which is the same for all points.

    (a) Verify this property.

    (b) Show also that the evolute of the equiangular spiral is again the equiangular spiral.

14. What is the envelope of the straight line solutions of the differential equation

$$y = x\frac{dy}{dx} + \left(\frac{dy}{dx}\right)^n ?$$

15. Show that the evolute of the involute of a curve $\mathcal{C}$ is $\mathcal{C}$ itself.

16. When does an involute of the evolute of a curve $\mathcal{C}$ coincide with $\mathcal{C}$? If the original curve $\mathcal{C}$ has no cusps then what information does Example 11-6h give about the involutes of the evolute?

17. In the text it was asserted that the solutions (25) of the system of differential equations (20) are all parametric representations of the same geometrical curve. Prove

    (a) any member of the family (25) can be obtained from any particular solution by rotation and translation;

    (b) given a solution of (20), any transformation of the solution by translation and rotation also is a solution.

18. The catenary (from Latin, catenarius, chain) is the curve assumed by a weighty chain or flexible cable of uniform density when it is hung between two support points. This is the shape of the cable between the towers of a suspension bridge before the deck is laid. The curvature function for the catenary is $k : s \rightarrow \kappa = \frac{1}{1 + s^2}$. Obtain the equation of a catenary and sketch the curve.

19. Let $\vec{X} = \vec{f}(s)$ represent a curve in three-dimensional space. For a space curve it is still true that $\frac{d\vec{t}}{ds}$ is perpendicular to the tangent $\vec{t} = \vec{f}'(s)$, and we define the principal normal $\vec{n}$ as the unit vector in the direction of $\frac{d\vec{t}}{ds}$. The curvature is now defined by $\kappa = \left|\frac{dt}{ds}\right|$ so that Equation (10) is still satisfied.

    (a) Obtain an expression for the curvature of a space curve in terms of any parameter, not necessarily arclength.

    (b) What is the curvature of a helix (Exercises 11-5, No. 6(f)) at any point?

    (c) Investigate whether Equation (11) must hold for a space curve.

Miscellaneous Exercises

1. (a) Show that the field of complex numbers is a vector space over the real numbers.

   (b) Show that the set of positive real numbers, $\mathcal{R}^+$, is a vector space over the field $\mathcal{R}$ of all real numbers where vector addition is defined as ordinary multiplication of real numbers (for p_1, $p_2 \in \mathcal{R}^+$ the vector sum is p_1p_2) and scalar multiplication is defined as exponentiation (for a scalar $\alpha \in \mathcal{R}$ and a vector $p \in \mathcal{R}^+$, the "product" of α with p is p^α).

2. (a) Draw the segments from the right angled vertex of a right triangle to the trisection points of the hypotenuse. Prove that the sum of the squares of the segments is proportional to the square of the hypotenuse and find the constant of proportionality.

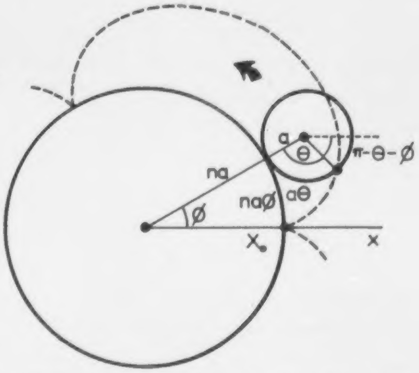
   (b) What is the constant of proportionality for the sum of the squares of the segments to the points of section of the hypotenuse into n equal parts?

3. Given that the sides of a triangle have lengths a, b, c, find the lengths of the medians.

4. (a) Prove that the cross product is not associative.

   (b) Under what conditions is the associative law for the cross product of three vectors satisfied?

5. The epicycloid of n cusps is the curve traced out by a point of a circle of radius a as it rolls in contact with and outside a fixed circle with radius na (see figure). The hypocycloid of n cusps ($n \geq 3$) is the curve traced out if the moving circle rolls on the inside of the fixed circle.



   (a) Obtain parametric equations for the epicycloid and the hypocycloid.

   (b) Prove that the epicycloid and hypocycloid of n cusps are simple closed curves.

   (c) Determine the areas enclosed by the epicycloid and hypocycloid of n cusps.

6. Consider a transformation of the plane in which the scales along the coordinates axes are changed independently:

$$(x,y) \longrightarrow (\xi,\eta)$$

where $\xi = ax$, $\eta = by$. Show that if κ is the curvature and θ the angle of inclination of a curve at a point then the transformed curve, at the corresponding point, has the curvature

$$\gamma = \frac{ab}{(a^2 \cos^2 \theta + b^2 \sin^2 \theta)^{3/2}} \kappa$$

7. Determine the radius of curvature of the evolute of $\mathcal{C}$ in terms of the radius of curvature of $\mathcal{C}$.

8. Find the envelope of the family of straight lines given by each criterion:

   (a) The product of the x- and y-intercepts is constant.

   (b) The sum of the x- and y-intercepts is constant c , where $c > 0$.

9. Obtain a parametric representation of the folium of Descartes given in Exercises 5-7, Number 13. Repeat that exercise in terms of the new representation.

10. (a) Prove the following generalization of the Law of the Mean. Let $\mathcal{C}$ be a plane curve given by $\vec{X} = \vec{r}(u)$ on $[a,b]$. If $\vec{r}$ is continuous on the closed interval $[a,b]$, if $\vec{r}\,'$ exists on the open interval (a,b) and is nowhere null, and if $\vec{r}(a) \neq \vec{r}(b)$, then there exists a tangent $\vec{t}_0 = \vec{r}\,'(u_0) \big/ |\vec{r}\,'(u_0)|$ for some u_0 in the open interval which is parallel to the chord joining the endpoints of the curve.

    (b) Express the Generalized Law of the Mean in terms of a coordinate representation of $\mathcal{C}$.

    (c) Prove or disprove the Generalized Law of the Mean for curves in E^3

11. (a) In Exercises 11-6, Number 13 we gave definitions for the principle normal $\vec{n}$ and curvature κ for a space curve $\vec{X} = \vec{r}(s)$. The vector $\frac{d\vec{n}}{ds}$ is perpendicular to $\vec{n}$ but it need not be parallel to $\vec{t}$. We introduce the binormal vector $b = t \times n$. Recall that

    (i) $$\frac{d\vec{t}}{ds} = \kappa\vec{n}$$

    and prove that there exists a scalar τ such that

    (ii) $$\frac{d\vec{n}}{ds} = -\kappa\vec{t} + \tau\vec{b}$$

    and

(iii) $$\frac{d\vec{b}}{ds} = -\tau\vec{n} \ .$$

The scalar τ is called the torsion of the curve. Equations (i), (ii), (iii) which generalize Formulas (10) and (11) of Section 11-6 are the Frenet-Serret equations for the curve.

(b) We have seen that if the curve is plane then $\tau = 0$. Prove, conversely, that if $\tau = 0$, then the curve is plane. (Hint: Show for given functions $\kappa = k(s)$, $\tau = \ell(s)$ that the solutions $\vec{r}$ of the Frenet-Serret equations subject to the initial conditions

(iv) $$\vec{r}(0) = \vec{x}_0 \ , \ \vec{r}\,'(0) = t_0$$

is unique.)

Chapter 12

MECHANICS

12-1. Introduction.

Mechanics has been, and still is, one of the continuing sources of new ideas for mathematics in general and calculus in particular. In fact, parts of the calculus were created in order to solve some of the mechanical problems we consider. Such scientists as Newton, Leibniz, Huyghens, the Bernoulli brothers, and Euler made fundamental contributions to both fields.

To a certain extent, the concepts and terminology of mechanics have entered the language, so that even without science courses you would be familiar with such ideas as mass, force, acceleration, energy, and work. The common meanings of these words need to be modified to agree with scientific uses. If we wish to present the foundations of mechanics carefully it is dangerous to build too heavily on intuition since in part intuition is not consistent with mechanics as we know it, and in part the subject has been built into our intuition by language and experience. On the other hand, a historical approach is not entirely successful since most of us do not know the extensive body of older scientific knowledge from which modern mechanics developed. Here, many of the concepts of mechanics are introduced as they originated historically, reformulated and used in a logical structure.

A typical problem in mechanics questions how objects move under given conditions, or when objects can be at rest with respect to each other. Fundamental in these questions are the assumptions that objects move or are at rest in something - physical space - and that it is possible to establish a measurement of time for which "before," "now," and "later" become quantitative statements. While the early workers in mechanics were quite aware of these questions, they were not greatly affected by them. Space was described by Euclid's synthetic geometry and later by Descartes' analytic geometry with the ordinary notion of distance. The concept of time was idealized as a quantity that could be measured as one measures distance along a line. These assumptions are not quite correct. It is important to appreciate the meaning of "correct" in this context. A mathematical statement is "correct" if it is obtained from axioms and definitions by the application of given rules of deduction. We speak of a physical idea as being "correct" if it is a fruitful way of thinking of phenomena that enables us to organize our information and predicts (or at least does not conflict with) new results. It is common, then, that older ideas are supplanted by newer ones as unexplained phenomena force us to investigate more deeply. When we say that euclidean geometry is not a correct description of physical space, we mean that we know of phenomena that are difficult to describe in euclidean terms and, further, that we have other geometrical systems which permit a simple description of these phenomena as well as others which are more familiar. Thus in more advanced works we are willing to sacrifice our familiar euclidean space in order to achieve elegance and unity in describing the physical world. While we readily admit the limitations of mechanics as we now describe it, its great success lies in its simplicity and its ability to give a fairly accurate description of systems ranging in size from many molecules to terrestrial objects to the solar system to stars, and often to galaxies.

In full awareness of its limitations, we adopt a description of physical space for the development of mechanics in terms of euclidean geometry. Points in space will be represented by position vectors $\vec{X}$; and it is assumed that we can measure lengths which correspond to euclidean distance $|\vec{X}|$. Implicit in these assumptions is the assertion that there is no distinguished origin or orientation of coordinate axes, so that no physical observation can be dependent on the choice of a coordinate frame. We also assume that we know how to measure time t and that we may describe the motion of a system in time by giving vector functions $\vec{r} : t \longrightarrow \vec{X}$ that describe the position of the particles of the system in a given space of vectors fixed in time. None of these assumptions is obvious or trivial, even though all are extremely natural to us. If you consider how you would verify the preceding statements, you will see that they cannot be separated from the remainder of our assumptions in mechanics, but that the subject must be verified as a whole. A proper justification is an exceedingly sophisticated matter. This descriptive part of mechanics is kinematics. In kinematics we seek as complete a description of the path $\vec{X} = \vec{r}(t)$ of a particle as possible.* Thus, kinematics is essentially the theory of curves introduced in Sections 11-5 and 11-6. There is certainly more to mechanics than kinematics, since we should like to be able to say when a given motion occurs. Dynamics is the part of mechanics that tells us what motion occurs in a given physical situation.

---

*The idea of a particle is an abstraction which also needs clarification. We may think of a particle as a bit of matter with dimensions so small with respect to the system of interest that it may be treated as though it is located at a point. Thus, even so extended an object as the earth may be treated as a particle if we are concerned only with its annual revolution about the sun, but not when we are concerned with its daily rotation on its axis. (However, even the latter motion can be treated like a particle motion in a theoretical space of suitably many dimensions, so that the concept takes on a different meaning from a more advanced viewpoint.) For us, the essential properties of a particle are its mass m and position $\vec{X} = \vec{r}(t)$. Associated with the position function $\vec{r}$ we have also the velocity $\vec{v} = \vec{r}'(t)$ and acceleration $\vec{a} = \vec{r}''(t)$.

We all have an idea of what it means to say that nothing is acting on or affecting an object. What happens to a body in motion, or at rest for that matter, if nothing acts on it? Quite consistently with our experience on the earth and with intuition Aristotle asserted that if an object in motion on or near the earth is not acted on, then it will come to rest. Galileo, by the study of motion on an inclined plane, obtained the law of inertia: in an ideal situation an object with nothing acting upon it would move in a straight line with constant velocity. Thus Galileo might say that if you roll a ball on the ground and it comes to rest, then something, the ground, produces a "force," which acts on the ball to bring it to rest. The great achievement of this and other idealizations is that they provide a basis for quantitative predictions of motion, as was not possible in the earlier physics. We adopt Galileo's assumption in a tentative form and we shall sharpen it considerably later. Thus, we say that an object under the influence of no other forces moves with constant velocity or remains at rest.

The effect of one object on another, a force, should at least correspond to our primitive notions of push and pull. As a result of Galileo's assumption we may say that a force produces a change in the velocity of an object or a force produces an acceleration of an object. We shall encounter many different kinds of forces later. Some of them are just pushes and pulls and we may apply them to one object or another. Others are more complicated and depend on both the object producing the force and the object experiencing it - and the force may even depend on the motion of the object experiencing it. In spite of the many complexities associated with the construct, force, it enables us to organize large areas of our experience. There is more to be said about force; at this point we say only that a force is that which produces an acceleration and it should include our intuitive idea of a push or pull.

Our ideas of motion with and without forces deserve a sharper look. We have assumed that accelerations are produced by forces. When we say that an object is accelerating, we can only mean it is accelerating as measured in a given coordinate frame. We might take a coordinate frame attached to the sun, or the earth, or a moving airplane, or a rotating phonograph record, or what you will.[*] The acceleration of a given object will be different in each of these systems, and hence we should have to say that there are different forces acting on the object in different coordinate systems. At the same time we might be misled by our intuitive feeling about a force as a push or pull into the idea that a force is a physical entity independent of coordinate system. But we have just observed that accelerations measured in different moving systems may not be the same.

At this point, we define an inertial coordinate system as one in which objects under the influence of no forces move with constant velocity. Our procedure is circular since we have already required a force to be something that produces a change in velocity. Consider a body accelerating in an inertial frame. Take another frame which moves with the body. With respect to the moving frame the body is at rest. Nonetheless the body is experiencing the force which made it move in the inertial frame. This makes sense only if we have a pre-established notion of force. Our ideas of forces arose from terrestrial observations and observations of the solar system. There is a natural coordinate system that we can use to describe such motions. We accept the assertions that the sun and planets move, but on the basis of simple observations we may say that the stars are approximately at rest with respect to each other. We know that the stars are in relative motion, but most observations over moderate periods of time show no such motion. Thus, all of classical mechanics refers to coordinate systems fixed with respect to the stars. Logically, what we are doing is using our observations in the special coordinate system of "fixed" stars to define what we accept as forces, i.e., as quantities that produce accelerations. We can then enlarge the class of coordinate systems we use to those in which an object under the influence of none of our known forces moves with uniform velocity. Thus, we may take the law of inertia stated above, Newton's First Law, as our first law of mechanics: an object under the influence of no force moves with constant velocity or remains at rest. Of course, every time we discover a new force we must check that it fits into this conceptual framework.

---

[*] The problem here is not the same as we encountered in our discussion of vectors where we obtained results independent of the orientation of coordinate axes. There, we considered different coordinate frames with arbitrary origins and orientations, but fixed in relation to each other; here the origin and orientation of coordinate frames may change relative to each other in time.

Though we may prefer to think of force as an entity not dependent on the coordinate system this is not possible. We cannot discuss motion and force independently of coordinate systems. The unsatisfactory nature of this restriction motivates a reformulation of physics without forces in a manner which does not distinguish special coordinate systems artificially. This was accomplished by Einstein in his theory of relativity.

We have raised very fundamental questions concerning the physical meaning of the concepts in classical mechanics, but we do not give these questions an entirely satisfactory resolution. We have only indicated the origin and the meaning of the concepts we use.

We now seek a quantitative relation between forces and acceleration. Without being too precise about the exact form of the relation between forces and accelerations, let us consider the results of some simple experiments. From our description of space, we know that acceleration, $\vec{a} = \frac{d^2\vec{X}}{dt^2}$, is a vector. A push or pull also has a magnitude and a direction; thus it is reasonable to suppose that force, too, is a vector. Further, for the simple forces that we know or can control, force and acceleration have the same direction. Finally, if we can apply separately two forces $\vec{f}_1$ and $\vec{f}_2$ to an object, then we may verify that the resultant acceleration when both forces are applied is the vector sum of the two separate accelerations. Thus we are led to guess that the force vector and acceleration vector are proportional, where the proportionality factor is not necessarily constant.

To learn about the proportionality factor between force and acceleration let us consider further experiments. Suppose we take a set of n equivalent lead cubes. The acceleration produced by applying a force to one of them is n times the acceleration produced by applying the same force to n of them rigidly bound together. Thus, the proportionality factor appears to be a measure of the amount of matter present. We call the amount of matter present in an object the mass of the object and we measure the mass by measuring the acceleration produced by a standard force on the object. Within the realm of nonrelativistic mechanics we assume that the mass of an object is fixed. In its preliminary form we take Newton's Second Law as

$$(1) \qquad \vec{F} = m\vec{a} = m\,\frac{d^2\vec{X}}{dt^2}\,,$$

where m is the mass of the object in question, $\vec{F}$ the vector sum of the forces applied to the object, and $\vec{a} = \vec{r}''(t)$ its acceleration.

Once we have stated this law we may describe more completely than before the logical structure of mechanics. We assume that we know how to measure space and time so that we may talk about acceleration. We work initially in the inertial coordinate system of the fixed stars and we use Newton's Second Law to measure mass by determining the acceleration produced on an object by a standard force. We may then study new forces in our inertial coordinate system by using Newton's Second Law again to measure the acceleration, and hence force produced. The procedure seems highly circular, for Newton's Second Law seems to be a definition of a force. The nontrivial part of the procedure is the assumption that a force is an independent entity. Once we have determined a force from (1) we may work with it, apply it to other objects, add it to other forces applied to bodies, and we still get valid results. Newton's Second Law can be used as a differential equation in which we give $\vec{F}$ and we solve for $\vec{r}$, but this is not its only use as we shall see.

At this point our form of Newton's Second Law is not quite adequate. The important concept of momentum was developed at the same time as the others previously discussed and we shall rephrase the second law in terms of momentum. The momentum of an object, a vector quantity usually denoted by $\vec{p}$, is the product of the mass times the (vector) velocity, $\vec{p} = m\vec{v} = m\vec{r}\,'(t)$. The importance of momentum is illustrated by the statement of a conservation law: the sum of the momenta of a system of objects is a constant provided the objects experience only forces exerted by other objects of the system. This law is, in fact, more fundamental than any of Newton's laws and is valid in the most advanced theories of physics. Ultimately, the law derives from a statement about the physical properties of space, but we are unprepared to discuss such questions. In the more advanced formulations neither mass nor velocity correspond to our usual ideas, but still it is possible to discuss a generalized momentum that is conserved. In terms of momentum, Newton's Second Law reads

$$\vec{F} = \frac{d\vec{p}}{dt}\,. \tag{2}$$

If we wish to discuss the motion of rockets (see Section 12-4iv) in which the mass of the "object" is not constant (matter is continuously ejected from the rocket), then (2) is a more convenient form of the second law than (1).

We may obtain Newton's Third Law from our statement of conservation of momentum together with the second law (conversely, Newton obtained conservation of momentum from the third law). Consider two objects and let the first exert a force $\vec{F}_{12}$ on the second, and let the second exert a force $\vec{F}_{21}$ on the first. Then

$$\vec{F}_{12} = \frac{d\vec{p}_2}{dt}$$

$$\vec{F}_{21} = \frac{d\vec{p}_1}{dt} .$$

Since $\vec{p}_1 + \vec{p}_2$ is a constant we find on adding the two equations $\vec{F}_{12} + \vec{F}_{21} = \vec{0}$, or the force one object exerts on a second is the negative of the force the second exerts on the first. In summary we take as Newton's laws in an inertial coordinate system,

1. an object under the influence of no forces moves with constant velocity;
2. if $\vec{F}$ is the vector sum of the forces applied to a body and $\vec{p}$ is its momentum, then

$$\vec{F} = \frac{d\vec{p}}{dt} ;$$

3. the force one object exerts on a second is the negative of the force the second exerts on the first.

We examine the implications of these laws in the remainder of the chapter.

Exercises 12-1

1. (a) Consider an inertial coordinate system, that is, a system in which Newton's laws hold. Let $\vec{r}(t)$ be the path of a particle in the given system and take new coordinates for which the particle path becomes $\vec{p}(t) = \vec{r}(t) + t\vec{v}$ where $\vec{v}$ is a constant vector. Describe what the change of coordinates means. Show that Newton's laws still hold provided forces are the same in both systems. This result is the Galilean Principle of Relativity.

   (b) Let $\vec{r}(t)$ be the particle path in an inertial system as in Part (a). Consider a new system in which the path of the particle is given by $\vec{p}(t) = \vec{r}(t) + \vec{q}(t)$. Show that the laws of motion in the new coordinate system are Newton's laws provided we add the inertial force $m\vec{q}''(t)$ to the total of the forces acting on each particle in the system.

   (c) What is the force experienced by an astronaut of mass m if the sole external force exerted upon him is the gravitational attraction mg of the earth and his rocket is accelerating upward with acceleration equal to $6g$?

12-2. Elementary Mechanical Problems.

In this section we examine a few elementary problems which we can solve directly with the use of Newton's laws and a few simplifying assumptions. These problems are interesting for themselves and also provide a basis for the general techniques of solution we develop in later sections.

(i) Motion of a projectile. We start with the motion of objects near the earth's surface under the influence of the earth's gravitational force. Coordinates fixed to the surface of the earth are not inertial because of the earth's rotation about its axis and its revolution around the sun as well as the sun's own motion relative to the fixed stars. Nonetheless, for the problems of motion near the earth's surface the corrections arising from the earth's motion are usually insignificant and we may ignore them. Only for large scale motions like that of a long range projectile is it necessary to account for the earth's motion. As a useful simplification, then, we assume that coordinates fixed to a point on the surface of the earth are inertial for motions near that point. Another useful idealization, Galileo's, is that the gravitational acceleration of any object is directed toward the center of the earth and has a constant magnitude independent of the object. According to Newton's Second Law, then, the earth exerts a force $m\vec{g}$ on a body of mass m, where $\vec{g}$ is directed toward the center of the earth and is constant in magnitude. We consider motions and displacements so small with respect to the size of the earth that we may simplify further and take the direction of $\vec{g}$ to be fixed also. For a first look at mechanics we need not worry about the net effect of these simplifications. It is sufficient for the present to know that there are errors, that they are relatively small, and that the errors may be estimated or a less crude model be used if the problem warrants it.

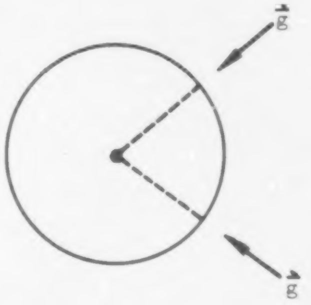


Figure 12-2a

Under the foregoing assumptions, for a particle moving solely under the influence of the earth's gravity, the path of the motion $\vec{X} = \vec{r}(t)$ satisfies Newton's Second Law, $m\frac{d^2\vec{X}}{dt^2} = m\vec{g}$, where $\vec{g}$ is constant. Consequently,

$$\vec{X} = \frac{t^2}{2}\vec{g} + t\vec{v}_0 + \vec{X}_0 . \tag{1}$$

(See Example 11-5d and Exercises 11-5, No. 8), where $\vec{v}_0$ is the velocity and $\vec{X}_0$ the position of the object at $t = 0$. The motion therefore occurs in a plane through the point $\vec{X}_0$ and parallel to the vectors $\vec{v}_0$ and $\vec{g}$. We now choose a cordinate frame with origin at $\vec{X}_0$, z-axis vertical and oriented downward, and x-axis in the plane of the motion, so that $\vec{g} = (0,0,g)$ where $g = |\vec{g}|$, and $\vec{v}_0 = (v_{0x},0,v_{0z})$. (See Exercises 12-2, No. 1.) In this coordinate frame, the path (1) has the representation $\vec{X} = (x,y,z)$, where

$$\begin{cases} x = v_{0x}\, t \\ y = 0 \\ z = \frac{1}{2} g t^2 + v_{0z}\, t . \end{cases} \tag{2}$$

If we eliminate t by putting $t = \frac{x}{v_{0x}}$ in the formula for z, we see that the path of the particle is a parabola. Moreover, the horizontal component of the velocity, the component parallel to the surface of the earth, is constant.

Now we solve the same problem by another method, a method which is seemingly more cumbersome but which is more generally useful. We wish to solve the differential equation

$$m\frac{d^2\vec{X}}{dt^2} = m\vec{g} . \tag{3}$$

If $m\vec{g}$ were a given function of time then we could solve (3) explicity by two successive integrations, as before. For many significant problems the forcing term does not depend on time, but on position alone. For such problems we adopt a different approach. Both methods can be illustrated by the given problem in which the forcing term mg is constant and, hence, can be considered as either a function of time or a function of position.

Note that

$$\frac{1}{2}\frac{d}{dt}(\vec{X}' \cdot \vec{X}') = \vec{X}'' \cdot \vec{X}' \tag{4}$$

so that by taking the dot product with $\vec{X}'$ in (3),

$$m\vec{X}'' \cdot \vec{X}' = m\vec{g} \cdot \vec{X}'$$

and then integrating once, we obtain

$$\frac{1}{2} m\tilde{X}' \cdot \tilde{X}' = m\tilde{g} \cdot \tilde{X} + c_1 . \tag{5}$$

In this way we have obtained a first order equation and achieved some simplification, but (5) is only one scalar condition on the three components of $\tilde{X}$ and it must be supplemented with two other conditions to define $\tilde{X}$. From the differential equation (3) itself we see that the components of the acceleration perpendicular to $\tilde{g}$ is zero, hence the components of velocity perpendicular to $\tilde{g}$ are constant. Thus (5) can be reduced to an equation for z alone:

$$\frac{1}{2} m\left(\frac{dz}{dt}\right)^2 = mgz + c_2 . \tag{6}$$

The quantity $\frac{1}{2} m\left(\frac{dz}{dt}\right)^2 - mgz$, which is constant, is called an integral of the motion. Sometimes an integral of the motion together with other information on the trajectory of an object will enable us to obtain a complete description of the motion. We leave to Exercises 12-2, Number 2, to show that (6) yields the same solution, (1).

The quantity $\frac{1}{2} mv^2$ in (6), where $v = |\tilde{X}'|$, is called the kinetic energy of the particle, and in this particular case the quantity $-mgz$ (or $-mgz$ plus any constant) is called the potential energy. Equation (6) tells us that the kinetic energy $\frac{1}{2} mv^2 = \frac{1}{2}(x'^2 + y'^2 + z'^2) = \frac{1}{2} mz'^2 + \text{const.}$ depends on position alone and not on time or the course of the preceding motion. When the total energy, kinetic plus potential in this case, is constant for a given motion, energy is said to be conserved. We shall return to these important concepts in Section 12-3.

In the foregoing solution we have assumed that the earth's gravity is the only acting force. The equations and their solution more truly describe motions near the surface of the moon, for on earth we cannot always disregard the action of air in retarding the motion. For moderate velocities a simple and experimentally acceptable assumption about the force of retardation is that it is proportional to the velocity and directed oppositely,

$$\tilde{F}_{ret} = -mk\tilde{v} , \qquad (k > 0) ,$$

where we have written the constant of proportionality in the form mk only to simplify the equations of motion.[*] From Newton's Second Law we obtain the equation of motion

$$\frac{d\vec{v}}{dt} = \vec{g} - k\vec{v} \, , \tag{7}$$

where $\vec{v} = \frac{d\vec{X}}{dt}$.

We will solve for $\vec{v}$ and then obtain $\vec{X}$. As we observed in Section 10-8 (p. 603), the solution of a linear equation such as (7b) may be written as any specific solution of the nonhomogeneous equation plus the general solution of the reduced equation. Since $\frac{\vec{g}}{k}$ is a particular solution of (7), we set

$$\vec{v} = \vec{u} + \frac{\vec{g}}{k} \, ,$$

where $\vec{u}$ satisfies the reduced equation

$$\frac{d\vec{u}}{dt} = -k\vec{u}$$

whose solution we know to be $\vec{u} = e^{-kt}\vec{u}_0$.

Thus

$$\vec{v} = \frac{\vec{g}}{k} + e^{-kt}\left(\vec{v}_0 - \frac{\vec{g}}{k}\right) \, ,$$

and on integration from 0 to t , we obtain

$$\vec{X} = t\,\frac{\vec{g}}{k} + \left(\frac{1 - e^{-kt}}{k}\right)\left(\vec{v}_0 - \frac{\vec{g}}{k}\right) + \vec{X}_0 \, . \tag{8}$$

We see for an object under the influence of gravity and atmospheric friction that there is an asymptotic velocity $\frac{\vec{g}}{k}$ independent of the initial conditions. The asympotic horizontal displacement is the horizontal component of $\frac{\vec{v}_0}{k}$ and thus depends only on the horizontal component of the initial velocity.

(ii) Oscillatory motion. In the solution of projectile problems we solved a problem with constant force $m\vec{g}$, and noted that if the force is a function of time only then the solution of Newton's equations of motion merely consists of direct integrations. For the interesting problems of mechanics the force is a function of position.

---

[*] The constant mk should not depend on the mass m , but only on the shape of the body and the smoothness of its surface.

We now study a number of such problems which serve as models of oscillatory and vibrational phenomena throughout physics and engineering. We begin with the simplest case.

Imagine a spring attached at one end to an immovable point of support and at the other to a particle of mass m compared to which the mass of the spring is negligible. The spring is said to be in equilibrium if it is at rest and no net force is acting upon it. Let the spring's axis be the x-axis and locate the origin at the free end of the spring in equilibrium. If the string is stretched or compressed by moving the particle along the x-axis, then it is found that the force acts to restore the spring to equilibrium and that it is proportional to the displacement from equilibrium. To a fair degree of accuracy, many systems behave like ideal springs, provided that they are not stretched too far; we shall see later why this is so. In any case, Hooke's Law states that the restoring force on a spring-mounted particle displaced from its equilibrium position 0 to $X = (x,0,0)$ is

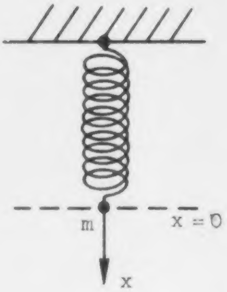


Figure 12-2b

(9)
$$\vec{F} = -k\vec{X} .$$

From Newton's Second Law, the equation of motion is

(10)
$$m \frac{d^2\vec{X}}{dt^2} = -k\vec{X} .$$

We consider only motion along the x-axis (however, see Exercises 12-2, No. 5), so that (10) becomes

(11)
$$\frac{d^2x}{dt^2} + \frac{k}{m} x = 0 , \qquad (k > 0) .$$

We immediately see (Section 10-8, Formula 12a), that the solution of this equation is a linear combination of sines and cosines:

(12)
$$x = x_0 \cos \sqrt{\frac{k}{m}}\, t + v_0 \sqrt{\frac{m}{k}} \sin \sqrt{\frac{k}{m}}\, t ,$$

where x_0 is the initial displacement and v_0 is the initial velocity. Thus, the motion is periodic of period $2\pi \sqrt{\frac{m}{k}}$. Note that we could also have obtained an integral of (11):

(13) $$\frac{1}{2}\, m\left(\frac{dx}{dt}\right)^2 + \frac{k}{2}\, x^2 = c$$

from which we could get the solution (12), (Exercises 12-2, No. 6). Energy is conserved if we call $\frac{k}{2}x^2$ the potential energy.

A mechanical system obeying an equation of the Form (11) is called a harmonic oscillator.

If a force acts to increase the displacement from equilibrium it is called a disturbing force. We leave as an exercise the solution of (11) when the force is a linear disturbing force $(k < 0)$ rather than a restoring force (Exercises 12-2, No. 7).

We reconsider the spring problem made realistic by the addition of the linear frictional force $-mr\frac{dx}{dt}$. With linear friction, Newton's Second Law yields

(14) $$\frac{d^2x}{dt^2} + r\frac{dx}{dt} + \frac{k}{m}x = 0\,, \qquad (r > 0)\,.$$

From Section 10-8(i) we recall the form of the solution of (14):
for $r^2 < 4\frac{k}{m}$,

(15) $$x = e^{-at}[x_0 \cos bt + \frac{1}{\omega}(v_0 + ax_0) \sin bt]\,,$$

where $a = -\frac{r}{2} > 0$ and $b = \sqrt{\frac{k}{m} - \frac{r^2}{4}}$:
for $r^2 > 4\frac{k}{m}$,

(16) $$x = c_1 e^{-\alpha t} + c_2\alpha^{-\beta t}\,,$$

where $\alpha = \frac{r}{2} + \sqrt{\frac{r^2}{4} - \frac{k}{m}} > 0$ and $\beta = \frac{r}{2} - \sqrt{\frac{r^2}{4} - \frac{k}{m}} > 0$. Thus, if $r^2 < 4\frac{k}{m}$ the motion is a damped oscillation, and if $r^2 > \frac{4k}{m}$, then after the first maximum displacement is reached the motion damps out monotonically. A cycle of the damped oscillation (15) is defined as the part of the motion between two successive maxima (or minima), the frequency as the number of cycles per unit time. The frequency is constant and equal to $\frac{b}{2\pi}$. The constant b is called the circular frequency of the system. Since b is a decreasing function of r, friction reduces the frequency of oscillation.

For many applications it is important to know what happens to a damped oscillator (15) when it is driven by an external force. A television set or radio contains many damped oscillators and the signal received and, later, the amplified signal act as external forces. An automobile in motion is another such system (the oscillators had better be damped, that's what shock absorbers are for) and the driving forces are exerted by the hot gases in the cylinders and the irregularities of the road. For most such systems we are primarily interested in the response to periodic driving forces.

We impose the periodic driving force $F \cos \omega t$ and ask what are the amplitude and frequency of the response. We wish then to study the solution of

$$(17) \qquad L[x] = \frac{d^2x}{dt^2} + r\frac{dx}{dt} + \frac{k}{m}x = F \cos \omega t .$$

Since we know the general solution of the reduced equation, we need only obtain any particular solution of (17). To get a particular solution of (17) we could use the Green's function technique of Section 10-8(ii), (see Exercises 12-2, No. 19), but in this case the work is simplified by the observation that $F \cos \omega t$ is itself the solution of a homogeneous linear differential equation with constant coefficients, namely $(D^2 + \omega^2)z = 0$. It follows that any solution x of (17) is also a solution of $ML[x] = 0$ where $M = D^2 + \omega^2$. Since M and L commute, it follows that if x_1 satisfies the equation $M[x_1] = 0$ and x_2 the equation $L[x_2] = 0$ then any linear combination $a_1x_1 + a_2x_2$ satisfies the equation $ML[a_1x_1 + a_2x_2] = 0$. (This is the same argument we used to derive Equation (8) of Section 10-8(i).) This result suggests that we attempt a solution of (17) of the form

$$(18) \qquad x_1 = A \cos (\omega t - \phi) .$$

We need not consider the term a_2x_2 since that is included in the general solution of the reduced equation for (17). Entering (18) in (17) we obtain

$$\begin{aligned} A[(\tfrac{k}{m} - \omega^2)\cos(\omega t - \phi) - r\omega \sin(\omega t - \phi)] &= A[(\tfrac{k}{m} - \omega^2)\cos\phi + r\omega\sin\phi]\cos\omega t \\ &\quad + A[(\tfrac{k}{m} - \omega^2)\sin\phi - r\omega\cos\phi]\sin\omega t \\ &= F\cos\omega t . \end{aligned}$$

Consequently (18) is a solution if the coefficients of $\cos \omega t$ and $\sin \omega t$ in this equation satisfy

$$A[(\frac{k}{m} - \omega^2)\cos \phi + r\omega \sin \phi] = F$$

$$(\frac{k}{m} - \omega^2)\sin \phi - r\omega \cos \phi = 0 .$$

From the second of these equations we obtain

(19a) $$\tan \phi = \frac{r\omega}{\frac{k}{m} - \omega^2} ,$$

and enter this result in the first equation and simplify to obtain

(19b) $$A = \frac{F}{\sqrt{(\frac{k}{m} - \omega^2)^2 + r^2\omega^2}} .$$

Thus we obtain the particular solution of (17) in the Form (18).

The general solution may now be written as the sum of the general solution (15) for the reduced equation and the particular solution (18):

(20) $$x = A \cos (\omega t - \phi) + Ce^{-at} \cos (bt - \psi)$$

where A and ϕ are fixed by (19) and C and ψ are fixed by the initial conditions. However, we see that the general solution decays away so that (18) is the asymptotic state which the solution approaches, no matter what initial conditions are imposed. We conclude that the "natural" frequency of oscillation given by b makes only a transient contribution to the frequency of the system, and that after a while the system oscillates with a circular frequency which differs insignificantly from the driving frequency ω . At the same time there is a "phase" lag τ in the oscillation of the system; that is, the peaks of the response lag behind the peaks of the driving force by the time $\tau = \frac{\phi}{\omega}$. The phase lag and the amplitude do depend on the natural frequency of the system as we see upon replacing r and $\frac{k}{2m}$ by the frequency $\frac{b}{2\pi}$ and decay constant a , through $r = -2a$, $\frac{k}{m} = a^2 + b^2$, to obtain

(21) $$A = \frac{F}{\sqrt{(a^2 + b^2 - \omega^2)^2 + 4a^2\omega^2}} , \quad \tan \phi = \frac{2a\omega}{a^2 + b^2 - \omega^2} .$$

Although the amplitude F of the driving force may not be large we see that it is possible for the amplitude A of the response to become extremely large for certain driving frequencies, particularly if the damping is slight (r small) and the spring is weak (k small). This phenomenon is known as resonance. In order to describe this phenomenon conveniently we set $\Omega = \frac{\omega}{\omega_0}$

and $c = \frac{r}{\omega_0}$ in (19b) where $\omega_0 = \sqrt{\frac{k}{m}}$ is the natural frequency of the undamped system (11), and replace A by

(22)
$$\alpha = \left(\frac{\omega_0^2}{F}\right) = \frac{1}{\sqrt{(1 - \Omega^2)^2 + c^2\,\Omega^2}} .$$

Now we plot α as a function of Ω for several given values of c, (Figure 12-3c)

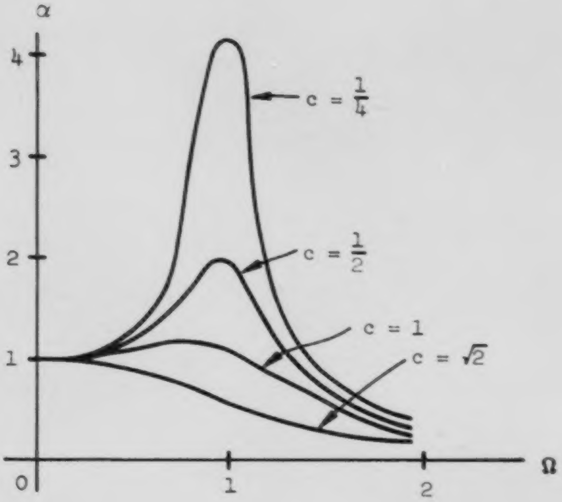


Figure 12-2c

For $c \geq \sqrt{2}$ there is no maximum for any positive driving frequency, but the least upper bound 1 is approached as Ω approaches 0 . For $c \leq \sqrt{2}$, there is a unique maximum of α corresponding to the resonant amplitude, $A_r = \frac{F}{\omega_0^2}\alpha_r$, where

(23)
$$\alpha_r = \frac{1}{\sqrt{c^2 - \frac{c^4}{4}}} ,$$

and this occurs at the resonant frequency $\frac{\omega_r}{2\pi}$ given by

(24)
$$\Omega_r = \frac{\omega_r}{\omega_0} = \sqrt{1 - \frac{1}{2}c^2} .$$

For $c < \sqrt{2}$ the maximum increases without bound as the damping parameter c approaches zero; at the same time the frequency corresponding to the maximum approaches the natural frequency of the undamped system. The curves of Figure 12-3c describe the tuning characteristics of elements in a radio or television receiver, and they also explain why a mechanical system vibrates intensely near the natural frequency and feebly elsewhere.

(iii) Motion of a charged particle. In addition to mechanical forces such as pushes and pulls and gravitational forces, classical physics is concerned with the forces of electricity and magnetism. A given particle has a charge q which is a measure of its responsiveness to electrical and magnetic influences. At each point X of space and moment in time t there is an electric vector $E = \vec{\mathcal{E}}(X,t)$ and a magnetic vector $\vec{B} = \mathcal{B}(X,t)$. The functions $\vec{\mathcal{E}}$ and $\vec{\mathcal{B}}$ are called, respectively, the electric and magnetic fields. The electromagnetic force on the particle of change q, called the Lorentz force, is

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}) \tag{25}$$

where $\vec{v}$ is the velocity of the particle. Accepting these as unexplained terms and unproven results (for us) from physics, we may yet investigate the motion of a charged particle in a given electromagnetic field under the force law (25).

The analysis of the trajectories of charged particles in electric and magnetic fields is essential in the physics of elementary particles (quantum electrodynamics): It is vital in the study of the interaction of the "solar wind" of charged particles emitted from the sun with the magnetic field of the earth and such of its effects as electrical storms, the aurora, and the Van Allen radiation belts. There are very few explicit solutions of the differential equations of particle motion in an electromagnetic field and much of our qualitative understanding of such motions comes from the simplest explicitly solvable case for which $\vec{E}$ and $\vec{B}$ are constant in time and space.

We consider the case of constant electromagnetic field in detail. Entering the Lorentz force in Newton's Second Law, we obtain the equation of motion

$$m \frac{d^2\vec{X}}{dt^2} = q(\vec{E} + \frac{d\vec{X}}{dt} \times \vec{B}) . \tag{26}$$

We assume $\vec{B} \neq \vec{0}$, otherwise the problem is analytically the same as (3). Since the force does not depend upon time we are encouraged to look for an energy integral; we take the dot product with $\frac{d\vec{X}}{dt}$ in (26), integrate, and find

$$\frac{1}{2} mv^2 = q\,\vec{E}\cdot\vec{X} + k \tag{27}$$

where $v = \left|\frac{d\vec{X}}{dt}\right|$ and k is constant. Consequently, if there is no electric field the kinetic energy is constant. (This conclusion holds even if the field is not constant.) Equation (27) is not sufficient information to permit the solution of (26); it is one scalar condition and we need three scalar conditions[*] to determine the three components of $\vec{X}$. We obtain another useful condition by taking the dot product with $\vec{B}$ in (26):

$$m\left(\frac{d^2\vec{X}}{dt^2}\cdot\vec{B}\right) = q(\vec{E}\cdot\vec{B})\ . \tag{28a}$$

This equation states that the component of the acceleration in the direction of $\vec{B}$ is constant and equal to $\frac{q}{m|\vec{B}|}(\vec{E}\cdot\vec{B})$. Let the subscript B indicate the component of any vector in the direction of $\vec{B}$, as in Exercises 11-4, Number 3. We integrate in (28a) to obtain the component of the motion in the direction of $\vec{B}$, namely,

$$\vec{X}_B = \frac{q}{m} t^2 \vec{E}_B + t\,\vec{v}_{0,B} + \vec{X}_{0,B}\ , \tag{28b}$$

where $\vec{X}_0$ and $\vec{v}_0$ are the initial position and velocity vectors.

Finally, we must determine the component of the motion perpendicular to $\vec{B}$. For $\vec{v}^B = \vec{v} - \vec{v}_B$, the component of velocity perpendicular to $\vec{B}$, we have the linear first order equation

$$m\frac{d\vec{v}^B}{dt} - q(\vec{v}^B \times \vec{B}) = q\vec{E}^B\ . \tag{29}$$

It would be possible to solve (29) by introducing coordinates immediately (Exercises 12-2, No. 19) but we shall continue with an invariant approach. To solve (29) we need only obtain the general solution of the reduced equation

$$m\frac{d\vec{u}}{dt} - q(\vec{u} \times \vec{B}) = 0\ , \tag{30}$$

where $\vec{u}\cdot\vec{B} = 0$, and add to $\vec{u}$ any particular solution of (29). Observe in (30) on taking the dot product with $\vec{u}$ that $\vec{u}\cdot\frac{d\vec{u}}{dt} = \frac{1}{2}\frac{d}{dt}(\vec{u}^2) = 0$. Consequently, $|\vec{u}|$ is constant. Since $\vec{u}$ has constant length, its orientation is given by the angle θ through which the direction of $\vec{u}$ has turned in the course of the motion from its initial direction $\vec{u}_0$. If we consider the motion as projected on a plane perpendicular to $\vec{B}$ it is geometrically evident from (30) that if $q > 0$ the sense of rotation as seen from the half space into which $\vec{B}$ points is clockwise (Exercises 12-2, No. 15). Now note that the Formulas (1) and (9) of Section 11-4 for the dot and cross products for the two vectors $\vec{u}$ and $\vec{u}_0$ remains valid if the angle θ is any angle (not necessarily restricted to the interval $0 \leq \theta \leq \pi$) measured positively in the counterclockwise sense of rotation from u to u_0 as viewed from the half space into which $\vec{B}$ points (Exercises 12-2, No. 16). Since $|\vec{u}| = |\vec{u}_0|$ we have on taking the cross product with $\vec{u}_0$ in (30),

---

[*] This method of counting conditions is an imprecise if convenient rule of thumb. In more complex situations the count may not be obvious.

$$\begin{aligned} m(\vec{u}_0 \times \frac{d\vec{u}}{dt}) &= m \frac{d}{dt} (\vec{u}_0 \times u) \\ &= m \frac{d}{dt} \{(|\vec{u}_0|^2 \sin \theta) \frac{\vec{B}}{|\vec{B}|}\} = m|\vec{u}_0|^2 \cos \theta \frac{d\theta}{dt} \frac{\vec{B}}{|\vec{B}|} \\ &= q[\vec{u}_0 \times (\vec{u} \times \vec{B})] \\ &= -q(\vec{u}_0 \cdot \vec{u})\vec{B} = -q|\vec{u}_0|^2 \cos \theta \vec{B} \end{aligned}$$

where we have used Formula (16) of Section 11-4 for the triple cross product. We conclude that

$$\frac{d\theta}{dt} = -\frac{q}{m} |\vec{B}|$$

thus the rotation rate $\omega = \frac{d\theta}{dt}$ is a constant, $\omega = -\frac{q}{m} |\vec{B}|$, and it is negative for $q > 0$ as asserted above. To complete the solution of the reduced equation we use $\theta = \omega t$, and write $\vec{u}$ in the form

$$(31) \qquad \vec{u} = (\cos \omega t)\vec{u}_0 - \frac{\sin \omega t}{|\vec{B}|}(\vec{u}_0 \times \vec{B}) .$$

Next we seek a particular solution of the nonhomogeneous equation (29). Observe that $\frac{d}{dt}$ commutes with the operator $L : \vec{v}^B \longrightarrow m \frac{d}{dt} \vec{v}^B + q(\vec{B} \times \vec{v}^B)$ in (29) and, directly generalizing the approach to the solution of (17), we seek a particular solution of (29) for which $\frac{d\vec{v}^B}{dt} = 0$, namely a constant vector. A constant solution of (29) must satisfy the condition

$$\vec{v}^B \times \vec{B} = -\vec{E}^B .$$

Since $\{\vec{v}^B, \vec{B}, -\vec{E}^B\}$ must therefore be a right-handed triple of mutually perpendicular vectors (assuming $\vec{F}^B \neq 0$), so also is $\{\vec{E}^B, \vec{B}, \vec{v}^B\}$. Therefore we may set

$$\vec{v}^B = k(\vec{E}^B \times \vec{B}) = k(\vec{E} \times \vec{B}) \ .$$

From this and the preceding equation it follows that $k = 1 = 1/|\vec{B}|^2$, and we obtain the particular solution of (29)

$$\vec{v}^B = \frac{d\vec{X}^B}{dt} = \frac{\vec{E} \times \vec{B}}{|\vec{B}|^2} \ . \tag{32}$$

From (31) and (32) we see that the component of the velocity perpendicular to $\vec{B}$ is the sum of the constant "drift" $\frac{\vec{E} \times \vec{B}}{|\vec{B}|^2}$ and a vector of constant length which rotates at a constant rate with circular frequency $\omega = \frac{q}{m} |\vec{B}|$. We leave to Exercises 12-2, Number 20 the proof that the motion perpendicular to $\vec{B}$ is given as the sum of a uniform straight line motion and a uniform circular motion. If $\vec{E} \cdot \vec{B} = \vec{0}$ the entire motion can be expressed as the sum of a uniform straight line motion and a uniform circular motion; the path of the particle in such a motion is called a trochoid. As a further exercise we leave the proof that if $\vec{E} = \vec{0}$ the path of the particle is a helix[*] (Exercises 12-2, No. 21).

(iv) Motion of a rocket. For the final problem of this section we consider a motion in which the mass of the "particle" changes in time. Such problems tend to be complex like the case of the falling raindrop which may lose water by evaporation and gain it by accretion and condensation. Here we consider the motion of a rocket propelled by the ejection of matter and under the influence of no external force.

Let M be the mass of the rocket, including fuel, m the mass of fuel ejected since the beginning of the motion, $\vec{v}$ the velocity of the rocket in an inertial frame, and $\vec{v}_e$ the velocity of the ejected matter in a frame fixed with respect to the rocket. Thus, in the inertial frame, the velocity of the ejected matter is $v + \vec{v}_e$. The force acting on the rocket is the reaction force of the escaping matter. This is calculated from the general form of Newton's Second Law, Equation (2) of Section 12-1, as the negative of the rate of change of momentum of the ejected matter. The change of momentum μ of ejected matter in the time interval $[t, t + \Delta t]$ is due entirely to the material ejected in that interval; hence, it is

---

[*] Defined in Exercises 11-5, No. 6(f).

$$\Delta\mu = [\frac{dm}{dt}(\vec{v} + \vec{v}_e) + \vec{\epsilon}]\Delta t ,$$

where $\lim\limits_{\Delta t \to 0} \vec{\epsilon} = \vec{0}$. (This requires the continuity of $\frac{dm}{dt}$ and $(\vec{v} + \vec{v}_e)$.)

Consequently, the reaction force $\vec{F}_r = -\frac{d\mu}{dt}$ on the rocket

(33) $$\vec{F}_r = -\frac{dm}{dt}(\vec{v} + \vec{v}_e) .$$

We assume that the propellent undergoes chemical, not nuclear, reactions so that the total mass of the system consisting of rocket and ejected matter remains constant:

$$M + m = M_0 .$$

We see, then, that

(34) $$\frac{dm}{dt} = -\frac{dM}{dt} .$$

From Newton's Second Law we have for the motion of the rocket

(35) $$M\frac{d\vec{v}}{dt} = \frac{dM}{dt}\vec{v}_e .$$

In order to solve this equation we must have some hypothesis about the rate of fuel consumption and the exhaust velocity (usually that both are constant), but special solutions are left to the exercises.

Finally, observe that near the surface of the earth we must add gravitational force so that the equation of motion becomes, with air resistance neglected,

(36) $$M\frac{d\vec{v}}{dt} = \frac{dM}{dt}\vec{v}_e + M\vec{g} .$$

In (35) we cannot consider $\vec{g}$ to be constant unless the range of the rocket is sufficiently restricted.

At this point we leave it to you to solve some of the problems of rocket motion (Exercises 12-2, Numbers 23 and 24).

Exercises 12-2

1. Show how to choose a fundamental set $\{\vec{i},\vec{j},\vec{k}\}$ for the derivation of (2) with the additional stipulation that $v_{0x} \geq 0$.

2. Show that Equation (6) yields the solution (1) of Equation (3).

3. (a) Show that in the limit of small air resistance (k approaches zero) that the solution (8) of (7) approaches the solution (1) of (3).

   (b) Shoot a particle upward; will it return to ground faster if encounters air resistance or no?

4. For velocities higher than those for which the derivation of (8) is valid, but lower than the speed of sound, it is found experimentally that the retarding force of the atmosphere is proportional to the square of the velocity,

$$\vec{F}_{ret} = -mk|\vec{v}|\vec{v} .$$

   (a) Determine the motion of a particle which moves in a vertical line under the influence only of gravity and air friction.

   (b) Re-examine question 3(b) for this form of air resistance.

5. (a) Solve the equation of motion (10) for a particle moving under the influence of a linear restoring force without restricting the motion to one dimension.

   (b) Show in this case that the path of the particle is an ellipse.

6. Find the Solution (12) of Equation (11) from the first integral of the Motion (13).

7. Solve Equation (11) when the force kx is a disturbing force $(k < 0)$ rather than a restoring force.

8. Use the Green's function technique of Section 10-8(ii) to obtain a particular solution of Equation (17).

9. (a) Find the general solution of (17) when $r^2 > 4\frac{k}{m}$, the case corresponding to the nonoscillatory damped solution (16) of the reduced equation.

   (b) Find the general solution of (17) when $r^2 = r\frac{k}{m}$, the so-called critically damped case.

10. (a) Which is nearer to the natural frequency of the undamped system $(r = 0)$ governed by (17), the natural frequency or the resonant frequency of the damped system?

(b) The "width" of the tuning curve $A = f(\omega)$ given by (19b) is a useful concept in broadcasting. If a receiver tuned to a station broadcasting at a given frequency has a sharply peaked tuning curve there will be no significant interference from stations broadcasting at nearby frequencies. A convenient measure of the width is

$$\frac{\omega^{+} - \omega^{-}}{\omega_r}$$

where ω^{-} and ω^{+} are respectively the frequencies below and above ω_r where the amplitude falls to value $\frac{\alpha_r}{\nu}$, where $\nu > 1$. Express this measure in terms of the constants of the system (17). Obtain an approximate representation for small c.

11. Obtain Formula (19b) with the aid of (19a) to complete the work indicated in the text.

12. What happens when you attempt to get a first integral of (14) by the method of multiplying by $v = \frac{dx}{dt}$? Consider the variation in time of the energy $E = \frac{mv^2}{2} + \frac{kx^2}{2}$? Is energy conserved?

13. Obtain the general solution for (17) when the applied frequency is equal to the resonant frequency for $r = 0$.

14. Observe for the undamped spring that the displacement x is an extremum when the velocity $v = 0$ and that the velocity v is an extremum when $x = 0$. Which of these statements is the more surprising?

15. In the text it is asserted as "geometrically evident" from (30) that if $q > 0$ the rotation of the direction of $\hat{u}$ is in the negative (clockwise) sense with respect to $\vec{B}$. If it is not evident to you, make it so.

16. For the derivation of (31), let ϕ be any angle measured positively in the counterclockwise sense of rotation from the direction $\hat{u}_0$ to the direction of $\hat{u}$ as seen from the half space into which $\vec{B}$ points. Show that ϕ may be taken as the angle in the definitions of dot product $\hat{u}_0 \cdot \hat{u}$ and cross product $\hat{u}_0 \times \hat{u}$ (1) and (9) of Section 11-4.

17. Verify Equation (32) by obtaining the result $k = \frac{1}{|\vec{B}|^2}$ given in the preceding text.

18. In the text we have merely solved (29) for the component of the velocity of the motion perpendicular to $\vec{B}$. Obtain the corresponding component of the displacement vector and give the complete solution of (26).

19. Solve (29) by introducing an appropriate coordinate frame.

20. Show that the component of particle motion perpendicular to the magnetic field B is the sum of a uniform straight line motion and a uniform circular motion.

21. Show that the motion of a particle in a constant electromagnetic field where $E = 0$ is a helix (ignore degenerate cases).

22. Discuss the motion of a particle under the influence of both a constant electromagnetic field and a constant gravitational field.

23. (a) Solve the equation of rocket motion (35) in one dimension under the assumption that the rate of fuel consumption $-\frac{dM}{dt}$ and exhaust speed $v_e = |\vec{v}_e|$ are constant.

    (b) For some purposes it is important that the acceleration not exceed some definite bound, for example, to limit the stress on an astronaut. Suppose the acceleration is set at this bound; replace the assumption in Part (a) by the assumption that the acceleration and v_e are constant and determine the way in which fuel should be consumed to achieve this result.

24. (a) Solve Equation (36) for the vertical ascent of a rocket in the gravitational field near the surface of the earth ($\vec{g}$ constant).

    (b) Consider the motion of Part (a) for a rocket at rest on the ground when $t = 0$. Find the relation between the fuel consumed to the velocity v. Estimate the fuel required to reach a given velocity assuming that it is by far the larger part of the initial mass M_0.

    (c) Determine the fuel consumption as a function of time under the same assumptions as Part (b) of Number 23.

12-3. Constraints. Use of Energy Conservation.

In Section 12-2 we examined problems of motion for which all the forces were stated explicitly. Many mechanical problems have geometrical side conditions on the path of a particle, for example, that the motion must follow a given curve or lie on a given surface. Such conditions are called constraints. A constraint generally implies the existence of a force of constraint which serves to keep the particle on its curve or surface. Since the velocity of the particle must be tangent to the curve or surface, the forces of constraint must act to prohibit motion perpendicular to the curve or surface and can be determined accordingly. Thus the normal forces on the object are not specified but may be found from the given constraint.

(i) Motion on an inclined plane. One of the simplest problems with constraints is to determine the motion of a particle on an inclined plane under the influence of a constant gravitational force. This problem was studied extensively by medieval scholars. Later, Galileo based many of his conclusions about mechanics upon results obtained with inclined planes. Consider a particle of mass m on a plane inclined to the horizontal at the angle θ, where $0 < \theta < \frac{\pi}{2}$ (Figure 12-3a). We introduce a coordinate system with z-axis along the intersection of the incline with a horizontal plane, y-axis perpendicular to the incline and directed upward, and x-axis upward along the incline. In this coordinate system the downward force of gravity is given by $m\vec{g} = (-mg \sin \theta, -mg \cos \theta, 0)$. If no other forces act, a particle initially at rest on the plane would fall straight down through the surface. If the particle is constrained to move on the incline, the plane must exert a force

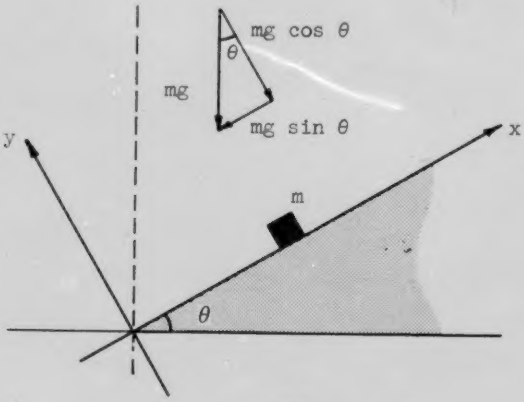


Figure 12-3a

to keep it there. In the given coordinate system, let (T_x,N,T_z) be the force exerted by the plane on the particle. Thus $\vec{T} = (T_x,0,T_z)$ is the component of force tangential to the plane and N is the normal component. From Newton's Second Law we obtain the differential equations of the motion

$$m\frac{d^2x}{dt^2} = -mg \sin\theta + T_x \tag{1a}$$

$$m\frac{d^2y}{dt^2} = -mg \cos\theta + N \tag{1b}$$

$$m\frac{d^2z}{dt^2} = T_z , \tag{1c}$$

and the Equations (1) are supplemented by the constraint

$$y = 0 , \qquad \text{for all } t . \tag{2a}$$

The constraint (2a) immediately determines the normal force exerted by the plane; namely,

$$N = mg \cos\theta . \tag{2b}$$

The tangential force $\vec{T}$ and the constraint are unrelated. In general, the tangential force is a frictional resistance and is given as an explicit function of the normal force and velocity. Without friction $(\vec{T} = \vec{0})$ a particle initially at rest slides down the incline with the constant acceleration $g \sin\theta$.

This section is devoted primarily to motion without friction, but we shall consider friction for motion on an inclined plane to show how it may be handled. A plausible assumption, which is supported by experiment if the velocity is not too large, is that the frictional resistance is proportional to the force N which holds the body against the plane. If μ is the constant of proportionality (the so-called coefficient of friction), the friction force is then

$$\vec{T} = -\mu N \frac{\vec{v}}{|\vec{v}|} \tag{3}$$

since friction acts to oppose the motion. For simplicity we confine our attention to motion along the x-axis. (The formulation of the more general problem which includes a z-component of velocity is left to Exercises 12-3, No. 1.) In that case,

$$T_x = -\mu N \operatorname{sgn} v = -\mu mg \cos\theta \operatorname{sgn} v ,$$

where $v = \frac{dx}{dt}$, and the equation of motion becomes

$$\frac{d^2x}{dt^2} = -g(\sin\theta + \mu\cos\theta \operatorname{sgn}\frac{dx}{dt}) .$$

There are two principal cases to consider: (a) if $\tan\theta > \mu$ then a particle with an initial downward velocity will continue down the incline with a constant acceleration of magnitude $\sin\theta - \mu\cos\theta$; if the initial velocity is upward there will be a retardation of magnitude $\sin\theta + \mu\cos\theta$ until the particle comes to rest, $v = 0$, then the particle will move downward with the absolute acceleration $\sin\theta - \mu\cos\theta$: (b) if $\tan\theta < \mu$ then the particle will slow down until it comes to rest, with a retardation of magnitude $\mu\cos\theta - \sin\theta$ if the initial velocity is downward, $\mu\cos\theta + \sin\theta$ if upward.

The preceding results are reasonably consistent with our experience of sliding objects. For an object at rest, however, the equation of motion is unrealistic. If $\tan\theta \leq \mu$ the particle will remain at rest, and not accelerate downward as the equation indicates. The effect of friction for a body at rest is to exactly oppose an applied force until it exceeds the critical value μN , after that, Equation (3) is applicable. Even this modification is not overly realistic. The angle θ_s at which a stationary object begins to slide is somewhat greater than one would predict from the coefficient of sliding friction. This fact is handled by introducing a coefficient of static friction $\mu_s = \tan\theta_s > \mu$ with the property that an object a rest will not be set into motion unless the tangential force exceeds $\mu_s|\hat{N}|$. Once the object is set into motion the frictional resistance is given by the coefficient of sliding friction. With this condition, we see that there are two possible courses for an initially upward motion with $\tan\theta > \mu$. If also $\tan\theta > \mu_s$ then the course of the motion is as described above, but if $\tan\theta < \mu_s$ then when the object comes to rest at the highest point of its ascent the object sticks to the plane and the motion stops.

It should be understood that even with the introduction of static friction our model is only an approximate idealization of actual frictional motions.

Consider now the somewhat more complicated motion of a particle which slides on the incline of a frictionless wedge which is itself free to slide without friction on a fixed horizontal plane (Figure 12-3b). We wish to describe the motion of both particle and wedge. For this purpose we take a fixed coordinate frame with x-axis horizontal and perpendicular to the edge of

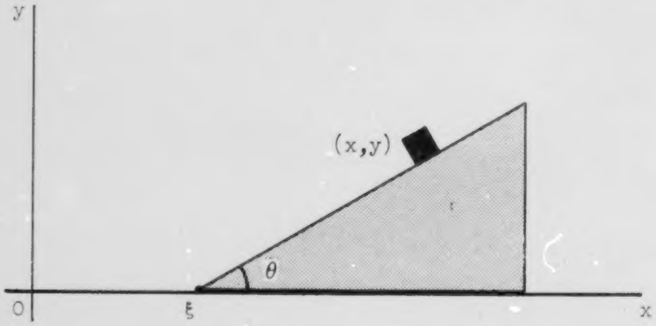


Figure 12-3b

the wedge and y-axis verticle. We assume that the initial velocities have no component perpendicular to the xy-plane so that the same is true of the entire subsequent motion. (See Exercises 12-3, No. 3.) Let the mass of the wedge be M and the position of the wedge be given by the location ξ of its edge on the horizontal plane. To the particle, assign the mass m and position (x,y). The constraint in this problem is

$$y = (x - \xi) \tan \theta . \tag{4}$$

Without friction, the force exerted on the particle by the wedge has no tangential component and can be written in the form $N(-\sin \theta , \cos \theta)$ where $N > 0$. By Newton's Third Law, the particle exerts the opposite force $N(\sin \theta , -\cos \theta)$ on the wedge. Newton's Second Law then yields for the particle

$$m \frac{d^2x}{dt^2} = -N \sin \theta \tag{5a}$$

$$m \frac{d^2y}{dt^2} = -mg + N \cos \theta , \tag{5b}$$

and for the wedge

$$m \frac{d^2\xi}{dt^2} = N \sin \theta . \tag{6}$$

We differentiate twice in (4) and use the result to eliminate $\frac{d^2y}{dt^2}$ from Equation (5b). Then we use (5) and (6) to eliminate $\frac{d^2\xi}{dt^2}$ and N and obtain a differential equation for x alone:

$$(7) \qquad \frac{d^2x}{dt^2} = -\frac{Mg \sin\theta \cos\theta}{M + m\sin^2\theta} .$$

Thus, the particle moves to the left with a constant acceleration. For the motion of the wedge we have

$$(8) \qquad \frac{d^2\xi}{dt^2} = \frac{mg \sin\theta \cos\theta}{M + m\sin^2\theta} ;$$

hence, as the particle moves to the left the wedge moves to the right. We need not obtain a differential equation for y, but solve (7) and (8) for x and ξ, and then obtain y from (4).

There are some other aspects of the motion which deserve attention. From (5a) and (6) we have

$$m\frac{d^2x}{dt^2} + M\frac{d^2\xi}{dt^2} = 0 ;$$

whence,

$$m\frac{dx}{dt} + M\frac{d\xi}{dt} = k .$$

We see then that the x-component of the momentum of the system particle plus wedge is constant. Note also that if the system is initially at rest then $k = 0$ and $mx + M\xi$ is constant in the motion. There is no external force acting in the x-direction and the kind of result we have just obtained is typical of mechanical systems without external and frictional forces.

(ii) Energy and its applications. A technique we have used several times is to obtain a first integral, a constant of the motion. Beginning with Newton's Second Law

$$(9) \qquad m\frac{d^2\vec{X}}{dt^2} = \vec{F}$$

we took the dot product with $\frac{d\vec{X}}{dt}$ and obtained

$$\frac{d}{dt}\frac{mv^2}{2} = \frac{d}{dt}\frac{1}{2}m\left(\frac{d\hat{X}}{dt}\cdot\frac{d\hat{X}}{dt}\right) = \hat{F}\cdot\frac{d\hat{X}}{dt}\,, \tag{10}$$

where $v = \left|\frac{d\hat{X}}{dt}\right|$. Then, if we found a function $V : \hat{X}\,,\ t \longrightarrow V(\hat{X},t)$ such that for $\hat{X} = \hat{r}(t)$

$$\frac{d}{dt}V(\hat{r}(t),t) = -\hat{F}\cdot\hat{r}'(t) \tag{11}$$

we integrated to obtain

$$\frac{1}{2}mv^2 + V(\hat{X},t) = k\,. \tag{12}$$

Although these steps appear to be nothing more than formal manipulations, a result of the Form (12) occurs so frequently that (12) and generalizations of (12) have the status of a basic physical principle, that energy is conserved (i.e., constant). It is not our purpose here to appropriately qualify and explain the general significance of this conservation. Let it be sufficient for now to show that the energy conservation principle (12) can be very useful in mechanical problems where it is shown to be valid.

We have already given a name to the term $\frac{1}{2}mv^2$ in (12), the kinetic energy. Set $T(t) = \frac{1}{2}mv^2$ and put (10) in the form

$$T(t_2) - T(t_1) = \int_{t_1}^{t_2}\hat{F}\cdot\hat{v}\,dt\,. \tag{13}$$

The integral $\int_{t_1}^{t_2}\hat{F}\cdot\hat{v}\,dt$ is called the work done on the object from time t_1 to time t_2. If the integral is positive, [illegible]ect of the work done is to increase the kinetic energy and hence t[illegible] speed of the motion; the ideas of work and energy do then have some relation [illegible]o the common meaning of these terms. If a function V satisfying (11) exists, then $V(\hat{X},t)$ is called the potential energy; the work done is then

$$V(\hat{r}(t_1),t_1) - V(\hat{r}(t_2),t_2)$$

so that (12) takes the form

$$T(t_2) + V(\hat{r}(t_2),t_2) = T(t_1) + V(\hat{r}(t_1),t_1)\,. \tag{14}$$

Thus if the potential energy decreases, the kinetic energy increases by the same amount.

Let us consider the effect of forces of constraint on the energy. Let us suppose we have a system with a constraint and an external force $\hat{F}$ which is derivable from a potential function V by (11). Further, we suppose the system to be frictionless so that the force exerted by the curve or surface on the particle consists entirely of the normal force of constraint $\hat{N}$. Newton's Second Law yields

$$m \frac{d^2\hat{X}}{dt^2} = \hat{F} + \hat{N} .$$

Since $\hat{v} = \frac{d\hat{X}}{dt}$ is a tangent vector to the curve or surface at $\hat{X}$, we have $\hat{N} \cdot \hat{v} = 0$. It follows that the force of constraint does no work and the energy conservation equation (12) is the same as that for the unconstrained system. For many motions under constraint we shall be able to obtain a complete description of the motion from (12) alone.

The first problem we treat by applying the energy principle is that of the motion of a pendulum. Consider a bob of mass m supported against gravity by a straight rigid rod of negligible weight which is attached to a fixed pivot about which it may freely rotate in a vertical plane (Figure 12-3c). We choose a coordinate frame in the plane of motion so that the z-axis is directed vertically upward and the x-axis to the right. Let θ be the angle the rod makes with the vertical. If ℓ is the length of the rod, the coordinates of the bob are given by $(x,z) = (\ell \sin\theta, -\ell\cos\theta)$. We assume that the system is frictionless so that the only force exerted by the rod is the force of constraint $\hat{N}$ which is directed normally to the path of the particle. Since $\hat{N}$ is perpendicular to the path it can be given in the form $\hat{N} = (-\lambda \sin\theta, \lambda\cos\theta)$. From Newton's Second Law we obtain the equations of motion

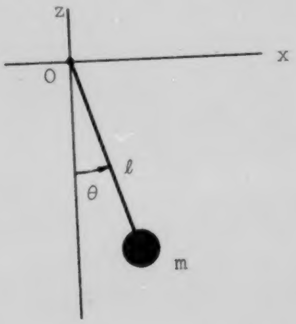


Figure 12-3c

(15a)
$$m\frac{d^2x}{dt^2} = -\lambda \sin\theta$$

(15b)
$$m\frac{d^2z}{dt^2} = -mg + \lambda\cos\theta$$

subject to the constraint

$$x^2 + z^2 = \ell^2 .$$

It may not be immediately obvious how to use the constraint to eliminate the normal force λ from (15) (see Exercises 12-3, No. 7). We shall proceed by calculating the energy integral instead. (We could use the result (12) directly, but choose to derive it again in this special case.) In (15a) we multiply by $\frac{dx}{dt}$, in (15b) by $\frac{dz}{dt}$, and add to get

$$\frac{d}{dt}\frac{m}{2}\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2\right] = -\frac{d}{dt}(mgz) - \lambda\left(\sin\theta\frac{dx}{dt} - \cos\theta\frac{dz}{dt}\right) .$$

We set $\frac{dx}{dt} = \frac{d}{dt}(\ell\sin\theta) = \ell\cos\theta\frac{d\theta}{dt}$, and $\frac{dz}{dt} = \frac{d}{dt}(-\ell\cos\theta) = \ell\sin\theta\frac{d\theta}{dt}$, in this equation and find

$$\frac{1}{2}\frac{d}{dt}\left[m\ell^2\left(\frac{d\theta}{dt}\right)^2\right] = -mg\ell\sin\theta\frac{d\theta}{dt} ;$$

whence, for $\omega = \frac{d\theta}{dt}$, the so-called angular velocity,

(16)
$$\omega^2 - \omega_0^2 = \frac{2g}{\ell}(\cos\theta - \cos\theta_0) ,$$

which is in the Form (14).

We cannot solve (16) explicitly for θ in terms of elementary functions but we can easily describe the general character of the motion. We do need Newton's laws to obtain one essential property of the motion; namely, what happens at a stationary point θ_0, a point where $\omega_0 = 0$, if such a point exists. Unless $\theta_0 = n\pi$, gravity has a non-zero component directed tangentially to the path; there is a net force on the particle, the pendulum is accelerating and θ must change. If $\theta_0 = n\pi$ corresponds to a stationary point the system is in equilibrium, at rest with no net force acting; later, we shall discuss equilibrium states in general.

The simplest case, that of no stationary points, occurs when the constant of integration (16), $\frac{g}{\ell}a = \omega_0^2 - \frac{g}{\ell}\cos\theta_0$ is sufficiently large, $a > 1$. We then have

$$\omega^2 = \frac{2g}{\ell}(a + \cos\theta) . \tag{17}$$

Since the angular velocity $\omega = \frac{d\theta}{dt}$ is never zero in this case, θ is a strongly monotone function of t . Thus, given enough initial speed, the course of the motion is a sequence of complete rotations of the pendulum around its pivot. If we fix the initial conditions so that $\theta = 0$ and $\omega > 0$ at $t = 0$, we obtain from (17)

$$t = \sqrt{\frac{\ell}{2g}} \int_0^{\theta} \frac{d\psi}{\sqrt{a + \cos\psi}} . \tag{18}$$

The motion is periodic with period

$$\tau = \sqrt{\frac{\ell}{2g}} \int_0^{2\pi} \frac{d\psi}{\sqrt{a + \cos\psi}} , \tag{19}$$

the time it takes the pendulum to make one complete turn on its pivot.

Next suppose $|a| < 1$ and again take the initial conditions $\theta = 0$, $\omega > 0$ at $t = 0$. Since $\omega > 0$ initially, the angular velocity remains positive until the stationary point

$$\theta_0 = \arccos(-a)$$

is reached at some time t_0 . During this interval the motion is still described by (18). At θ_0 the acceleration is directed back down along the path, ω changes sign and θ decreases until the next stationary point is reached. In this time interval the motion is then given by

$$t - t_0 = -\sqrt{\frac{\ell}{2g}} \int_{\theta_0}^{\theta} \frac{d\psi}{\sqrt{\cos\psi - \cos\theta_0}} . \tag{20a}$$

the next stationary point is reached at $\theta_1 = -\arccos(-a) = -\theta_0$ when $t = 3t_0$ and then the motion is reversed again. The motion is periodic, the period being

$$\tau = \sqrt{\frac{2\ell}{g}} \int_{-\theta_0}^{\theta_0} \frac{d\psi}{\sqrt{\cos\psi - \cos\theta_0}} . \tag{20b}$$

The function $t \longrightarrow \theta$ behaves much like the cosine function of the same period. Like cos, the complete function can be described by the first quarter cycle as defined by (18) for $\theta_0 \leq \theta \leq 0$.

We may easily see that, for small oscillations, a cosine function is in fact a good approximation to the solution. Differentiate in (16) to obtain

$$2\boldsymbol{\omega}\frac{d\boldsymbol{\omega}}{dt} = 2\,\frac{d\theta}{dt}\,\frac{d^2\theta}{dt^2} = \frac{-2g}{\ell}\,\sin\theta\,\frac{d\theta}{dt}\ ,$$

whence,

$$\frac{d^2\theta}{dt^2} = -\frac{g}{\ell}\sin\theta\ . \tag{21}$$

For small amplitudes we know that $\sin\theta$ is a very good approximation to θ; in fact, if $|\theta| < \epsilon \le \frac{\pi}{2}$, we have from Example 7-5b

$$|\theta - \sin\theta| \le \frac{\epsilon^3}{6}\ . \tag{22}$$

We then seek an approximation of the nonlinear equation (21) by a solution of the linear equation

$$\frac{d^2\theta}{dt^2} = -\frac{g}{\ell}\,\theta\ , \tag{23}$$

which is the equation of a harmonic oscillator (see Section 12-2(ii)). Equation (23) has the solution

$$\theta = A\sqrt{\frac{g}{\ell}}\cos\left(t\sqrt{\frac{g}{\ell}} + \phi\right)\ . \tag{24a}$$

Under the initial conditions $\theta = \theta_0$, $\boldsymbol{\omega} = 0$ at $t = 0$, we obtain

$$\theta = \theta_0\cos t\sqrt{\frac{g}{\ell}}\ . \tag{24b}$$

Thus the approximate solution with amplitude θ_0 has the period $2\pi\sqrt{\frac{\ell}{g}}$. From (22) with $\epsilon = \theta_0$ we can obtain an estimate of the difference between the exact and the approximate solutions and the difference in their periods (Exercises 12-3, No. 9).

(iii) Motion of a particle constrained to move on curve. Consider the frictionless motion of a particle constrained to move on a curve $\mathbf{X} = \mathbf{q}(s)$ where s is arclength along the curve and the external force is given by a potential $V(\mathbf{X})$ which does not depend upon time. From the equation of energy conservation (12)

$$\frac{m}{2}\left(\frac{d\hat{X}}{dt}\right)^2 = \frac{m}{2}\left[\frac{d\hat{X}}{ds}\,\frac{ds}{dt}\right]^2 = k - V(\hat{q}(s)) \ ,$$

or, since $\left|\frac{d\hat{X}}{ds}\right| = 1$,

$$\frac{m}{2}\left(\frac{ds}{dt}\right)^2 = k - U(s) \tag{25}$$

for $U : s \longrightarrow V(\hat{q}(s))$. In particular, when V is a constant function, the net external force is zero, and from (25) the particle moves with constant speed $\frac{ds}{dt}$ along the curve, a result which nicely supplements Newton's First Law.

Example 12-3. The cycloidal pendulum

Although the motion of a pendulum is approximately sinusoidal for small amplitudes with a period independent of amplitude, there is actually a dependence of period upon amplitude which connot be neglected for large amplitudes (see Exercises 12-3, No. 10). In order to build an accurate clock based on the motion of a pendulum is therefore necessary either to accurately control the energy imparted to the pendulum and so control the amplitude, or, somehow, to modify a pendulum so that the period is independent of the amplitude. Huyghens[*] found an extremely clever solution of the second kind, the cycloidal pendulum.

Consider a particle moving frictionlessly on the cycloid

$$\begin{cases} x = a(\phi - \sin\phi) \\ z = -a(1 - \cos\phi) \end{cases} , \qquad (0 < \phi < 2\pi) \ ,$$

under the influence of gravity, with coordinates chosen as for the pendulum (Figure 12-3d). Thus the external force has only a z-component, $F_z = -mg$, and it is derivable from the external potential $V(\hat{X}) = mgz$. With ϕ as the parameter instead of arclength, the energy equation (25) becomes

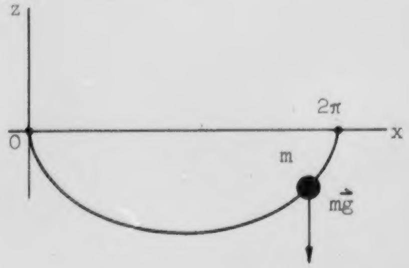


Figure 12-3d

---

[*] Huyghens, C., Dutch mathematician and physicist, 1629-1695.

(26) $$\frac{m}{2}\left(\frac{ds}{d\phi}\right)^2\left(\frac{d\phi}{dt}\right)^2 = a\,mg(1 - \cos\phi) + K.$$

In this equation put

$$\left(\frac{ds}{d\phi}\right)^2 = \left(\frac{dx}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2 = 4a^2\sin^2\frac{\phi}{2}$$

(compare the solution of Exercises 11-5, No. 6(b)), to get, with a new constant, k ,

(27) $$\sin^2\frac{\phi}{2}\left(\frac{d\phi}{dt}\right)^2 = \frac{g}{a}\left(\sin^2\frac{\phi}{2} - k\right) .$$

To keep the motion within the interval $0 < \phi < 2\pi$, we take $0 < k < 1$. We may then draw the same sort of conclusions as for the oscillation of a pendulum. The motion consists of an oscillation between the turning points $\phi_0 = 2\arcsin\sqrt{k}$ and $\phi_1 = 2\pi - \phi_C$. With the initial condition $\phi = \phi_0$, $\frac{d\phi}{dt} = 0$ at $t = 0$ we obtain from (27) for the times before ϕ_1 is reached,

$$t = \sqrt{\frac{a}{g}}\int_{\phi_0}^{\phi}\frac{\sin\frac{1}{2}\psi}{\sqrt{\sin^2\frac{1}{2}\psi - \sin^2\frac{1}{2}\phi_0}}\,d\psi$$

$$= \sqrt{\frac{a}{g}}\int_{\phi_0}^{\phi}\frac{\sin\frac{1}{2}\psi}{\sqrt{\cos^2\frac{1}{2}\phi_0 - \cos^2\frac{1}{2}\psi}}\,d\psi ;$$

whence,

$$t = 2\sqrt{\frac{a}{g}}\left[\frac{\pi}{2} - \arcsin\left(\frac{\cos\frac{1}{2}\phi}{\cos\frac{1}{2}\phi_0}\right)\right] .$$

We enter $\phi = 0$ in this equation to obtain the quarter period, and, so obtain for the full period,

(28) $$\tau = 4\pi\sqrt{\frac{a}{g}} .$$

Observe that the period of the oscillation is independent of the amplitude given by ϕ_0 .

As we have seen there is a great deal we can learn about a curvilinear motion which satisfies conservation of energy even when we cannot integrate (25). The methods we have applied are special but may be used whenever the potential function U is known. Consider the graph $y = U(s)$ and draw the line $y = k$ (Figure 12-3e). In the figure, the graph of U lies below the line only when $a < s < b$ or $c < s$.

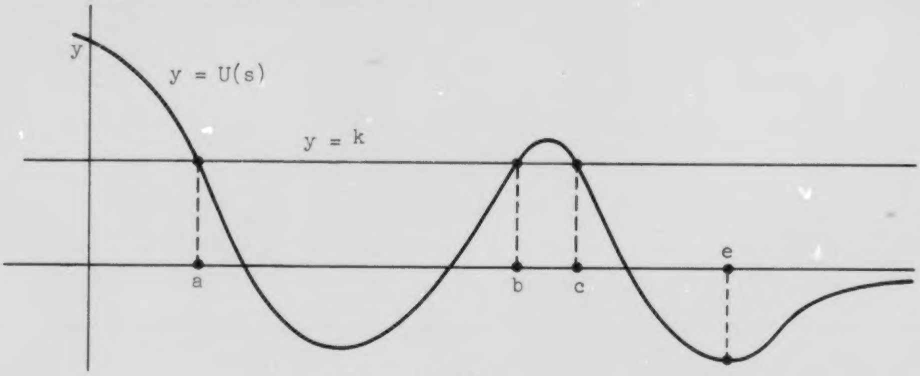


Figure 12-3e

For a given total energy k , motion can only take place in one of the intervals $a \leq s \leq b$ or $c \leq s$. Furthermore, as $U(s)$ increases, the speed $\left|\frac{ds}{dt}\right|$ decreases. A particle slows down as it approaches any of the points a , b or c , and it will reach such a point at zero speed. We know from the definition of potential function (11) that $-\vec{F}\cdot\frac{d\vec{X}}{dt} = \frac{d}{dt}V(\vec{q}(s))) = \frac{d}{dt}U(s(t))$; hence on multiplying by $\frac{dt}{ds}$, that

$$\vec{F}\cdot\frac{d\vec{X}}{ds} = -U'(s) \quad . \tag{29}$$

If the speed vanishes at a point where $U'(s) \neq 0$ then the net force on the particle, the tangential component of $\vec{F}$, does not vanish; consequently the particle is accelerated along the curve in the direction of decreasing potential. Physicists call a trough in the graph of U like that on the interval $[a,b]$ a "potential well" and think of it as a trap for low energy particles.

To be precise we say there is a potential well on an interval $[a,b]$ where $U(a) = U(b) = k$, provided derivatives $U'(a)$ and $U'(b)$ are non-zero and $U(a) < k$ for s in the open interval (a,b) . We leave it as an exercise to show that a particle in the potential well, beginning from rest at a , takes a finite time to go from a to b and an equal time to come back (Exercises 12-3, No. 11). The motion is periodic, and under the initial condition $s = a$, $\frac{ds}{dt} = 0$, the first half-cycle of the motion is given by the improper integral

(30) $$t = \sqrt{\frac{m}{2}} \int_a^s \frac{d\sigma}{\sqrt{U(a) - U(\sigma)}} \qquad (s \in [a,b]) ,$$

and the period is

(31) $$\tau = \sqrt{2m} \int_a^b \frac{d\sigma}{\sqrt{U(a) - U(\sigma)}} .$$

We leave it as a further exercise to show that if $U(a) = k$ corresponds to a maximum of the potential, with $U'(a) = 0$ and $U''(a) < 0$, then it takes infinite time for a particle of total energy k to move from a neighboring point s to a . In other words, the improper integral does not converge.

Finally there are unbounded motions like those which occur for $s \geq c$. A particle with total energy k proceeding to the left from e , for instance, will reach the point c , turn around, and move out to the right boundlessly.

If U has a minimum at s_0 we may under simple conditions approximate U in a neighborhood of s_0 by a quadratic polynomial which has the same value and first two derivatives at s_0 as U . (This generalizes the tangent approximation of Section 5-7. Such approximations will be studied in detail in Chapter 13 under the heading of Taylor's Theorem). Let $s_0 = 0$ and take as the approximating polynomial $W(s) = \alpha + \beta s^2$ where $\alpha = U(0)$ and $\beta = \frac{1}{2} U''(0)$. We suppose $\beta > 0$ so that the minimum for both U and W at $s = 0$ is a strong one. A particle in the potential well of $W(s)$ satisfies the differential equation $\frac{d^2s}{dt} + \frac{2}{m} s = 0$ and thus behaves like a simple harmonic oscillator. Insofar as the approximation to U by W is applicable, the motion of a particle restricted to a sufficiently small neighborhood of the minimum is approximately that of a simple harmonic oscillator. Mathematically the validity of the approximation of U by W is justified by proving continuous dependence of the solution of (25) on the given function U . We do not pursue the question of continuous dependence here. We only wish to indicate why an enormous variety of mechanical systems behave like simple harmonic oscillators in the neighborhood of their equibria.

From the graph of the potential we can also immediately see whether a particle in equilibrium, that is, at rest with no net force acting upon it, will remain near equilibrium if disturbed slightly. Consider the potential of Figure 12-3f. Since equilibria can occur only at points where the net force

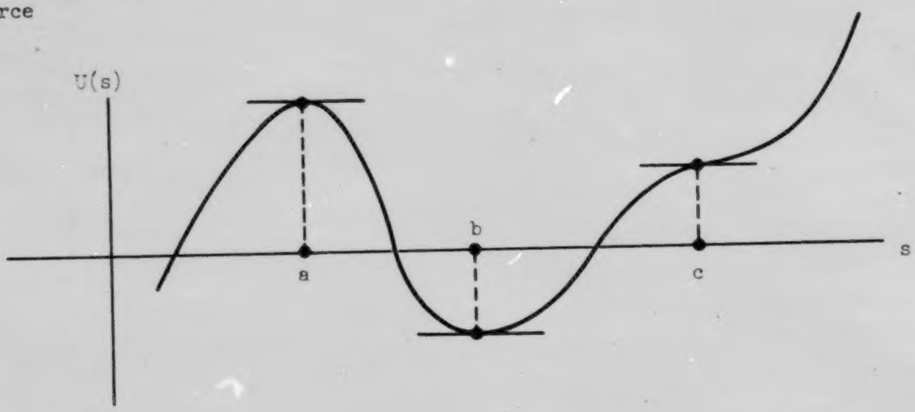


Figure 12-3f

vanishes, they can occur only where $\frac{dU}{ds} = 0$, in this case at $s = a$, b , c . A perturbation of a particle at rest at one of these points consists of a slight change in its position or velocity which results in a slight change in its energy. Clearly a perturbation of a particle at rest at either a or c results in an eventual large displacement from the point in question. Such an equilibrium is called unstable in the sense that any perturbation, no matter how small, will initiate a motion which takes the particle far from its original equilibrium. A particle in equilibrium at b , on the other hand, will stay near the bottom of its potential well if it is disturbed slightly; such an equilibrium is called stable. Thus a strong local minimum corresponds to a stable equilibrium, a strong local maximum to an unstable one.

It should be emphasized that a stable equilibrium need not be "very" stable. For the two parts of Figure 12-3g we see that even though an equilibrium is a strong local minimum, or even an absolute minimum, to boot, a "small" disturbance may push the particle far from equilibrium. However, it remains true that there is a minimum increment of energy, though it may be "small" which is required for such a gross displacement.

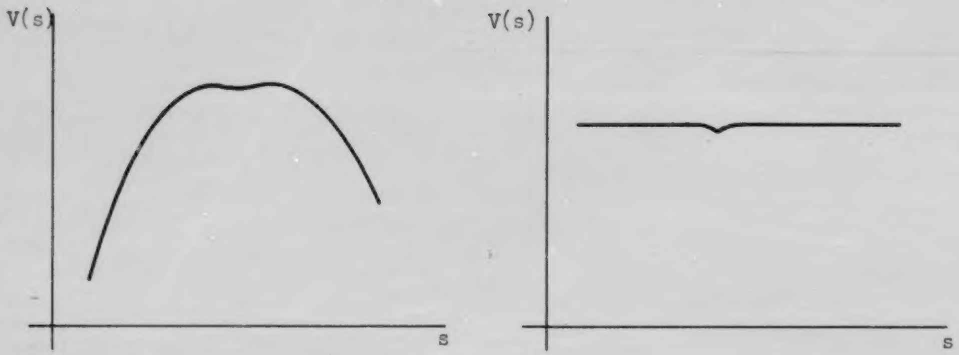


Figure 12-3g

Stability questions in particle mechanics and other fields are usually among the most subtle and difficult problems encountered in the applications. With the continuing refinement of our science and technology the issue of stability has become a central issue in several areas. For example, in the study of servomechanisms, systems which can respond to their own actions, uncontrolled instabilities are to be avoided since they can produce extreme destructive oscillations akin to resonances. The current effort to control the hydrogen fusion reaction resorts to magnetic fields to confine and stably contain the reaction; both the theory and practice of achieving stability in this area taxes our present day resources to their limits.

Exercises 12-3

1. Obtain the Newtonian equations of motion for a particle moving on an inclined plane subject to gravity and frictional forces of the Form (3). Do not assume the z-component of velocity is zero as in the text.

2. Obtain the complete equations of motion for the system consisting of a particle constrained to move on a frictionless wedge which slides on a horizontal plane. Verify that the motion is two-dimensional if the initial velocity is perpendicular to the edge of the wedge.

3. What is the normal force N exerted by the particle on the frictionless wedge?

4. Consider the motion of a particle sliding without friction on a wedge when the wedge slides with the coefficient of friction μ against the horizontal plane.

   (a) Obtain the equations of motion corresponding to (7) and (8) under the assumption that $\frac{d\xi}{dt} > 0$. (Hint: Consider the equation of motion for the y-component of position for the wedge so that the normal force exerted by the plane on the wedge may be taken into account.

   (b) Give that the system is initially at rest, under what conditions will the wedge be set into motion? (Ignore the difference between static and sliding friction).

   (c) Determine the order by size of the normal forces exerted by the particle sliding frictionlessly on a wedge for the three cases, stationary wedge, wedge sliding frictionlessly on the horizontal plane, wedge sliding with friction.

5. Obtain the energy conservation principle for the system consisting of the particle sliding on a frictionless wedge, Equations (5) - (6). (Hint: Take as the kinetic energy of the system the sum of the kinetic energies for particle and wedge.)

6. What is the magnitude λ of the force of constraint for the pendulum?

7. The text states that it is not immediately obvious how to use the constraint to eliminate the constraint force λ from the equations of motion (15a, b). Show how to do it.

8. Describe the motion of the pendulum when $\theta = \pi$ is a stationary point, that is, when $a = 1$ in (18).

9. Estimate the difference between the error of approximation to the amplitude of the exact solution (20a) of the pendulum problem by that of the approximate solution (24a).

10. (a) Show that the period of the pendulum as given by (20) is an increasing function of θ_0 .

11. Show for a particle oscillating in a potential well that the motion is periodic. Show further that the time to traverse the well from one side to the other is equal to the time to come back.

12. Let $s = 0$ correspond to a maximum of the potential U , with $U(0) = k$, $U'(0) = 0$, and $U''(0) < 0$, where U'' is continuous in a neighborhood of 0 . Show that a particle in the neighborhood with total energy k and velocity directed toward 0 takes infinite time to reach 0 .

13. Consider a particle of mass m which slides frictionlessly on a vertical circular hoop, where the hoop itself is spinning about its vertical diameter with constant angular speed ω . Describe the motion. (Hint: Use the energy conservation law in the Form (25) where s is arclength on the hoop).

14. Consider a particle moving on a curve $\vec{X} = \vec{r}(s)$ subject to the conservation law (25) with a potential of the form

$$U(s) = A + Bs^2 + s^3F(s)$$

where $B > 0$ and the derivative F' exists and is bounded. Take $W(s) = A + Bs^2$ as an approximation to $U(s)$ near $s = 0$. Let the arclength for the motion under the potential U be given by $s = \phi(t)$, under the potential W , by $\sigma = \psi(t)$. Show for small amplitude oscillations that σ and s are close together for the half-cycle beginning with the initial states $\phi(0) = \psi(0) = \alpha > 0$, $\phi'(0) = \psi'(0) = 0$

12-4. Angular Momentum and Central Forces.

The mathematical effect of a constraint is to reduce the number of independent position coordinates in the solution of a mechanical problem. For example, by constraining the motion of a particle to a curve in Section 12-3, we reduced the number of essential position coordinates to one, arclength along the curve, which could then be used as a parameter to give the position of the point. Sometimes the constraints were tacit rather than explicit, as in those simple cases where we assumed, not proved, the motion to be confined to a point or line parallel to an external force. In all these cases, the effect was to reduce the solution of the equations of motion to a one-dimensional problem, the solution of a differential equation for one position coordinate. In this section we treat other problems reducible to one dimensional motion.

(1) Central force fields. We treat first central force fields in which the force on a particle at a point X is collinear with the position vector $\vec{X}$. In particular we consider special central forces of the form $\vec{F} = \phi(\rho)\vec{X}$ where $\rho = |\vec{X}|$. Thus in some coordinate system the magnitude of the force at X depends only on the distance to X from O and is parallel to the ray OX. In this situation Newton's Second Law yields

$$m \frac{d^2\vec{X}}{dt^2} = \phi(\rho)\vec{X} \quad (1)$$

If we take the cross product with $\vec{X}$ in (1) we obtain a result which does not involve the force explicitly:

$$m\vec{X} \times \frac{d^2\vec{X}}{dt^2} = m \frac{d}{dt}\left(\vec{X} \times \frac{d\vec{X}}{dt}\right) = \vec{0} \quad (2)$$

From this we obtain an integral of the motion,

$$m\vec{X} \times \frac{d\vec{X}}{dt} = m\vec{K} \quad (3)$$

where $\vec{K}$ is a constant vector (compare Example 11-5c). The angular momentum $m\vec{X} \times \frac{d\vec{X}}{dt}$ is another basic mechanical concept. Equation (3) asserts that in a central force field angular momentum is conserved.

The constant vector $\vec{K}$ in (3) is determined by the initial position and velocity. If we take the dot product in (3) with $\vec{X}$ we obtain

$$\vec{X} \cdot \vec{K} = 0$$

so that the entire trajectory lies in the plane through the origin and perpendicular t , $\vec{K}$, that is, the plane containing the line OX_0 and parallel to the vector $\vec{V}_0$ where $\vec{X}_0$ is the initial position, and $\vec{V}_0$, the initial velocity, of the vector. With this knowledge the original three dimensional problem has been reduced to two dimensions. If it should happen that $\vec{K} = 0$ the trajectory lies on a straight line and the problem is further reduced to one dimension; the proof is left to Exercises 12-4, Number 1. We shall assume $\vec{K} \neq 0$ henceforth

Consider the area in the plane of motion swept over by the segment $\overrightarrow{OX}$ during the time interval from t_1 to t_2 (Figure 12-4a). With a prime indicating the derivative with respect to t we write

$$(4) \quad A = \int_{t_1}^{t_2} \vec{k}\cdot(\vec{X}\times\vec{X}')dt = |\vec{K}|(t_2 - t_1)$$

where $\vec{k} = \dfrac{\vec{K}}{|\vec{K}|}$, and A is the area given by the integral (7) of Section 11-6. (Note that the straight segments $\overline{OX_1}$ and $\overline{OX_2}$ contribute zero to the integral around the closed curve since the tangent and position vector are collinear along these segments.) Equation (4) states the principle of conservation of momentum in the form of one of Kepler's laws* (the so-called second law): the position vector to a particle from the origin of a central force field sweeps out equal areas in equal times. Kepler discovered this law for planetary motions about the sun.

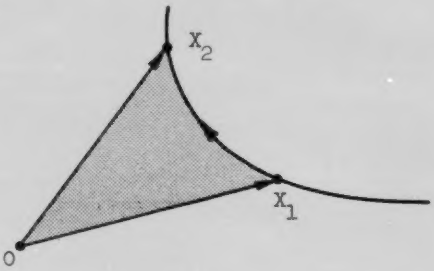


Figure 12-4a

It seems a pity to toss off in a line one of the three laws of planetary motion which cost Kepler more than two decades of arduous computation and manipulation of the data of planetary motion, yet such is the power of the calculus applied to Newton's laws (Newton inherited the first two of these from Galileo). Consider, however, that the earlier work of Kepler, Galileo, and others was necessary to the development of mechanics in Newton's hands; Newton is supposed to have said, "If I have seen a little farther than others it is because I have stood on the shoulders of giants." Oddly, although he obtained his results by the calculus, Newton, in his great Mathematical Principles of Natural Philosophy* presented his work geometrically to make it acceptable to his contemporaries.

---

*Kepler, J., German 1571- 1630.

*Philosophiae Naturalis Principia Mathematica (1687).

It is convenient to rewrite Kepler's law (4) in polar coordinates within the plane of the trajectory. If the polar coordinates of the particle are ρ and θ, then $\vec{X} = (\rho \cos \theta, \rho \sin \theta, 0)$ and $\vec{X} \times \vec{X}' = (0,0,\rho^2\theta')$. With this, (3) can be written in the form

$$\rho^2\theta' = \frac{dA}{dt} = K \tag{5}$$

where $\vec{K} = (0,0,K)$. Since $K > 0$ by (3) and (4), θ is a strongly increasing function of t. Thus the sense of rotation of $\overrightarrow{OX}$ about the origin is constant.

Until now we have used only the property that the force at $\vec{X}$ is directed along the line OX through the origin. The preceding results are valid even if the dependence $\phi(\rho)$ on radial distance in (1) is replaced by a more general kind of dependence $\phi(\vec{X},t)$ on both position and time. Now we make use of the explicit form of the force given in (1) to obtain an energy conservation law. We show that a force of the type appearing in (1) is derivable from a potential. From the identity

$$\frac{d\rho}{dt} = \frac{d}{dt}|\vec{X}| = \frac{\vec{X}\cdot\vec{X}'}{\rho},$$

(see Exercises 11-5, No. 4a), and the relation $\vec{F} = \phi(\rho)\vec{X}$, it follows that

$$\vec{F}\cdot\vec{X}' = \phi(\rho)\frac{d\rho}{dt}$$

so that the force is the time derivative of a potential $V(\rho)$; namely,

$$-\vec{F}\cdot\vec{X}' = \frac{d}{dt}V(\rho),$$

where

$$V(\rho) = -\int_{\rho_0}^{\rho} \phi(r)r\,dr. \tag{6}$$

The energy conservation law (12) of Section 12-3 states that

$$\frac{m}{2}|\vec{X}'|^2 + V(\rho) = E$$

where the total energy E is constant. In terms of the polar representation of $\vec{X}$ this equation becomes

$$\frac{m}{2}[\rho'^2 + \rho^2\theta'^2] + V(\rho) = E.$$

From this equation we may eliminate θ' by (5) to obtain

$$\frac{m}{2}\rho'^2 + V(\rho) + \frac{mK^2}{2\rho^2} = E \;. \tag{7}$$

This is the equation of a one-dimensional motion with the potential $V(\rho) + \frac{mK^2}{2\rho^2}$ for which we may use the graphical techniques of Section 12-3.

When we are interested mainly in the trajectory, not the time dependence of the motion, we may use θ as the parameter instead of t. Since θ is a strongly monotone function of t, the two parameters are equivalent, thus we are spared the complications of the pendulum problem attributable to the stationary points where the motion reverses direction. From (5) and (7) we find

$$\frac{m}{2}\frac{K^2}{\rho^4}\left(\frac{d\rho}{d\theta}\right)^2 + V(\rho) + \frac{mK^2}{2\rho^2} = E \;. \tag{8}$$

(ii) Inverse square forces. Planetary motion. Kepler stated two other laws of planetary motion. Kepler's First Law states that the orbit of a planet is an ellipse with the sun at one focus. Newton postulated that the sun exerted a central force on the planets and found that this law of Kepler's implied that the magnitude of the force of attraction of the sun is proportional to the "inverse" square of the distance between the planet and the sun.* From this conclusion to proceed on to the universal law of gravitation is a lesser step. The historical evidence indicates that Newton had thought out basic ideas of the calculus and mechanics together with the universal law of gravitation in his early twenties.

We shall not follow Newton in obtaining the inverse square force from Kepler's law. Instead, we shall assume the inverse square law and verify that the trajectories given by Kepler's First Law occur under its influence (but see Miscellaneous Exercises, No. 2). Specifically, we assume the central force

$$\vec{F} = -\frac{\alpha}{\rho^2}\frac{\vec{X}}{|\vec{X}|} = -\frac{\alpha}{\rho^3}\vec{X} \tag{9a}$$

with proportionality constant α, and show that any ellipse with focus at the origin is among the possible trajectories under this law. In general, Equations (7) and (8) can be used to find the trajectories for any central force law. We shall leave as an exercise the problem of obtaining the trajectories under an inverse square law from either (7) or (8), (Exercises 12-4, No. 2). Here, we use special tricks, particularly appropriate to the inverse square law, for obtaining these trajectories.

---

*The word "inverse" is to be taken in the sense of reciprocal.

From Newton's Second Law,

(9b)
$$m\vec{X}'' = -\frac{\alpha}{\rho^3}\vec{X}$$

we have on taking the cross product with $\vec{K}$,

$$\frac{d}{dt}\, m(\vec{K} \times \vec{X}') = -\frac{\alpha}{\rho^3}\vec{K} \times \vec{X} .$$

From (5) we have $\vec{K} = (0,0,\rho^2\theta') = \rho^2\theta'\vec{k}$ and we rewrite the preceding equation in the form

(10)
$$\frac{d}{dt}\, m(\vec{K} \times \vec{X}') = -\alpha\theta' \frac{\vec{k} \times \vec{X}}{\rho} .$$

Now we write the right side of (10) as a time derivative. From

$$\begin{aligned} \frac{\vec{k} \times \vec{X}}{\rho}\frac{d\theta}{dt} &= (-\sin\theta , \cos\theta , 0)\frac{d\theta}{dt} \\ &= \frac{d}{dt}(\cos\theta , \sin\theta , 0) \\ &= \frac{d}{dt}\frac{\vec{X}}{\rho} \end{aligned}$$

we obtain (10) in the form

$$\frac{d}{dt}\, m(\vec{K} \times \vec{X}') = -\frac{d}{dt}\frac{\vec{X}}{\rho} ;$$

whence

(11)
$$m\vec{K} \times \vec{X}' = -\alpha\frac{\vec{X}}{\rho} - \vec{B} .$$

Since $\vec{X}$ and $\vec{K} \times \vec{X}'$ are parallel to the plane of the trajectory, we conclude that so also is the constant vector $\vec{B}$. Choose coordinates in the plane so that $\vec{B} = (\beta,0,0)$ with $\beta > 0$. Take the dot product with $\vec{X}$ in (11) and use

$$(\vec{K} \times \vec{X}') \cdot \vec{X} = -\vec{K} \cdot (\vec{X} \times \vec{X}') = -K^2$$

(see Exercises 11-4, No. 15) to get

(12)
$$\rho = \frac{mK^2}{\alpha + \beta\cos\theta} .$$

Equation (12) is the polar representation of a conic section which has a focus at the origin and x-axis as an axis of symmetry. The trajectory is an ellipse if $\alpha^2 > \beta^2$, a parabola if $\alpha^2 = \beta^2$, and a hyperbola if $\alpha^2 < \beta^2$.

The general character of the trajectory can be determined at any instant by the distance from the origin and the speed. To verify this observation, put (11) in the form

$$\vec{B} = -\alpha \frac{\vec{X}}{\rho} - m\vec{K} \times \vec{X}'$$

to obtain

$$\beta^2 = |\vec{B}|^2 = \alpha^2 + \frac{2\alpha m}{\rho} \vec{X} \cdot (\vec{K} \times \vec{X}') + m^2 |\vec{K} \times \vec{X}'|^2 \ ;$$

whence,

$$\beta^2 - \alpha^2 = m^2 K^2 v^2 - \frac{2\alpha m K^2}{\rho} \ . \tag{13}$$

where $v = |\vec{X}'|$. It follows that the trajectory is an ellipse parabola or hyperbola according to whether $m^2K^2v^2 - \frac{2\alpha mK^2}{\rho}$ is negative, zero, or positive.

The escape velocity is the smallest speed necessary for an unbounded trajectory, a parabola or hyperbola. Thus the escape velocity is $\sqrt{\frac{2\alpha}{m\rho}}$, provided $\alpha > 0$, that is, provided the central force $\vec{F}$ given by (9a) is an attraction. If the force is a repulsion, $\alpha < 0$, only hyperbolic orbits exist.

We study the elliptic trajectory in detail. With $\beta > 0$, this case occurs for $\alpha > \beta$. Since the focus lies on the major axis of symmetry the orbit has the general appearance of Figure 12-4b where O , the center of the field of force is at one focus of the ellipse, the other focus is indicated by O' , and the center of the ellipse by C . The semi-major axis of the ellipse is given by (12) as $a = \frac{1}{2}(\rho_0 + \rho_\pi)$ where ρ_0 is the value of ρ at $\theta = 0$, and ρ_π at $\theta = \pi$; thus

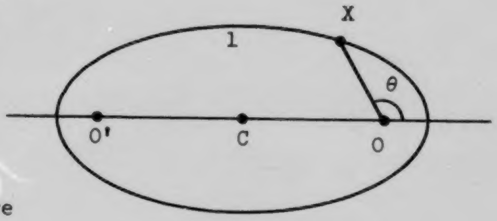


Figure 12-4b

$$a = \frac{mk^2\alpha}{\alpha^2 - \beta^2} \ . \tag{14}$$

We may eliminate β from (13) by means of (14) to obtain the equation

$$v^2 = \frac{2\alpha}{m\rho} - \frac{\alpha}{am} \tag{15}$$

which gives the speed at any point of the ellipse in terms of position on the ellipse and the constants m , a , α . Next we prove Kepler's Third Law, the square of the period is proportional to the cube of the semiaxis of the orbit. From Equation (4), the period τ is proportional to the area of the closed orbit,

$$\tau = \frac{\pi ab}{K} ,$$

where b is the semi-minor axis of the ellipse which is given, with the help of (14) by

$$b = \frac{mK^2}{\sqrt{\alpha^2 - \beta^2}} = K\sqrt{a}\ \sqrt{\frac{m}{\alpha}} ,$$

(see Exercises 12-4, No. 3). We then find for the period

$$\tau = \pi a^{3/2} \sqrt{\frac{m}{\alpha}} . \tag{16}$$

If the sun were fixed and exerted the sole force on the planets, an inverse square attraction, then Kepler's laws would constitute a complete description of planetary motion. Luckily for Kepler, the data of his time were not so precise as to reveal the slight discrepancies from his laws. Given Kepler's laws, Newton framed his law of gravitation:[1] given two particles with masses m_1 and m_2 and positions $\hat{X}_1$ and $\hat{X}_2$, respectively, the force exerted by the object of mass m_2 on the object of mass m_1 is

$$- \frac{Gm_2m_1(\hat{X}_2 - \hat{X}_1)}{|\hat{X}_2 - \hat{X}_1|^3} \tag{17}$$

where G is a constant of the universe. In this it is assumed, even though the masses of the sun and planets are not located at points, that these bodies can be treated as particles. By means of an integration it is possible to show for a sphere whose density is a function of radius alone that its gravitational attraction on an object outside its surface is the same as that of a particle with the mass of the sphere, and located at its center. Such integrations can be performed very simply with techniques for integrating functions of several variables which are developed a little later in the calculus. It appears that Newton delayed twenty years until he could satisfy himself of the validity of this proposition before he announced the law of gravitation. [2] (It is, in fact, not hard to show for any shape for distances from the body which are large in comparison to its size, that the departures from the inverse square law of attraction are negligible. But remember that Newton, in order to prove such results, had to develop the necessary calculus first.)

---

[1] Good enough to serve for the physics of more than two and one half centuries between Newton and Einstein.

In the inertial frame of the fixed stars, the equations of motion of the sun with mass indicated by m_1 and position by $\bar{X}_1$ and a planet with mass m_2 and position $\bar{X}_2$ are, by (17),

$$m_1\bar{X}_1'' = -\frac{Gm_1m_2(\bar{X}_2 - \bar{X}_1)}{|\bar{X}_2 - \bar{X}_1|^3} \tag{18a}$$

and

$$m_2\bar{X}_2'' = \frac{Gm_1m_2(\bar{X}_2 - \bar{X}_1)}{|\bar{X}_2 - \bar{X}_1|^3} . \tag{18b}$$

Upon adding the two equations (18) we obtain an equation exhibiting conservation of momentum,

$$\frac{d}{dt}(m_1\bar{X}_1' + m_2\bar{X}_2') = 0 . \tag{19}$$

The point $\bar{C} = \frac{m_1\bar{X}_1 + m_2\bar{X}_2}{m_1 + m_2}$ is called the center of mass of the system. From (19), $\bar{C}'' = \bar{0}$, thus the center of mass has zero acceleration. To obtain the equation of motion in terms of the position vector $\bar{X} = \bar{X}_2 - \bar{X}_1$ of the planet with respect to the sun, divide in (18a) by m_1 in (18b) by m_2 and subtract to find

$$\bar{X}'' = -\frac{G(m_1 + m_2)\bar{X}}{\rho^3} \tag{20}$$

where $\rho = |\bar{X}|$, as before. This equation has precisely the same form as (9b) with $\alpha = G(M + m)m$ and all our earlier results hold with this value. Consequently, in a coordinate frame fixed with respect to the sun Kepler's first laws still hold; the orbit of a planet is an ellipse with the sun at a focus, the vector $\bar{X}$ sweeps out equal areas in equal times, and the period τ of the motion is given by

---

[2] Cajori, F. in Sir Isaac Newton, pp. 127-190 (History of Science Society, Balitmore, 1928).

$$\tau = \frac{\pi a^{3/2}}{\sqrt{G(m_1 + m_2)}} . \tag{21}$$

Since m_1 is much greater than m_2 for all the planets, the approximation $\alpha \approx GMm$ is quite good. However, for a massive planet like Jupiter, for which $m_2 \approx .001\, m_1$ the difference from the accurate period (21) is easily detectible.

The problem we have solved treates two bodies in isolation. The general problem of finding the trajectories of three gravitating bodies is already far deeper. The best we have been able to do is to find computational methods for approximating the solution in particular cases and even that may be very complex. For the general n-body system we may expect the problem to be even less tractable. For the solar system, however, the effect of any other planet on the orbit of a given planet about the sun can be considered as a small perturbation of the basic two-body motion. In this way practical approximate solutions of the equations of motion for the planetary system are obtained. There are other interesting and difficult questions which one may naturally ask, the most important being the question of stability of the solar system. Can the mutual interactions between the planets grossly affect the orbit of any one of them? Reassuringly, for the ideal particle picture we have been using, Laplace* in 1773 showed that no great disruption of the solar system can occur through internal interactions, a magnificant achievement.

In their efforts to solve the problem of celestial mechanics, mathematicians from Newton's day to Gauss created the calculus and powerful methods of analysis which are useful for much else. In our times mathematics stimulates itself as much as it has been stimulated by its applications. A mathematician today can pursue his career without knowledge of any one of the significant applications of mathematics. Still, like the mythical giant Antaeus, mathematics gains strength by contact with the mother Earth and is made vigorous by meeting the problems imposed by other human concerns.

(iii) The spherical pendulum. Complete conservation of angular momentum implies that the force is a central force (Miscellaneous Exercises, No. 2(a)). However, there are some problems with noncentral forces for which some components of the angular momentum remain constant in the motion. We conclude with another example in which such partial conservation of angular momentum is used to reduce the problem to one dimension. We consider a pendulum but allow the bob to move on a sphere instead of constraining the motion to a circle in a vertical plane. We have the same equation of motion as before, namely

---

*Laplace, P. S., French, 1749 - 1827.

$$m\vec{X}'' = m\vec{g} - \lambda\vec{X} \tag{22}$$

where $\vec{X}$ is subject to the constraint $|\vec{X}|^2 = \ell^2$ and the constraint force $-\lambda\vec{X}$ is directed toward the support of the pendulum. The energy integral is

$$\frac{1}{2}\, m|\vec{X}'|^2 - m\vec{g}\cdot\vec{X} = mk \ . \tag{23}$$

This is an equation in the three components of $\vec{X}$. The constraint permits us to eliminate one component. In order to eliminate another component and reduce the problem to one dimension we need another condition which does not involve the unknown factor λ. We eliminate λ from (22) by taking the cross product with $\vec{X}$, and find the derivative of the angular momentum

$$m\,\frac{d}{dt}(\vec{X}\times\vec{X}') = m\vec{X}\times\vec{g} \ . \tag{24}$$

Thus angular momentum is not conserved. Note, however, that the component of the derivative in the direction of $\vec{g}$ is zero. Therefore we have conservation of the component of angular momentum parallel to $\vec{g}$; namely,

$$\frac{m}{g}[\vec{g}\cdot(\vec{X}\times\vec{X}')] = m\mu \ , \tag{25}$$

where μ is constant.

We choose a coordinate frame which simplifies the elimination of two coordinates from Equations (23) and (25) and the constraint condition. The z-axis is taken vertically upward so that $\vec{g} = (0,0,-g)$. We use polar coordinates r , θ with $x = r\cos\theta$, $y = r\sin\theta$. The constraint $|\vec{X}|^2 = x^2 + y^2 + z^2 = \ell^2$ can then be written in the form

$$r^2 + z^2 = \ell^2 \ . \tag{26}$$

From (25), we find

$$r^2\theta' = \mu \ , \tag{27}$$

and from (23),

$$\frac{1}{2}[r'^2 + r^2\theta'^2 + z'^2] + gz = k \ . \tag{28}$$

If $\mu = 0$ the problem reduces to the circular pendulum treated in Section 12-3 as you may verify (Exercises 12-4, No. 6). We assume hereafter that $\mu \neq 0$. (From (27) it then follows that θ is a strongly monotone function of t.) With (27) we eliminate θ from (28) and with (29) we eliminate r to obtain the one differential equation for z,

$$\frac{\ell^2}{\ell^2 - z^2} z'^2 + \frac{\mu^2}{\ell^2 - z^2} + 2gz = 2k$$

or

$$z'^2 + \frac{2(gz - k)(\ell^2 - z^2)}{\ell^2} = -\frac{\mu^2}{\ell^2} . \tag{29}$$

Equation (29) corresponds to a one-dimensional motion with the "potential" $\frac{2(gz - k)(\ell^2 - z^2)}{\ell^2}$ and "total energy" $-\frac{\mu^2}{\ell^2}$.

We analyze the motion qualitatively, as in Section 12-3, by studying the potential as a function of z. The potential function has three real roots at $z = \pm \ell$, $z = \frac{k}{g}$. We distinguish three cases $\frac{k}{g} < -\ell$, $-\ell < \frac{k}{g} < \ell$ and $\ell < \frac{k}{g}$, (Figure 12-4c).

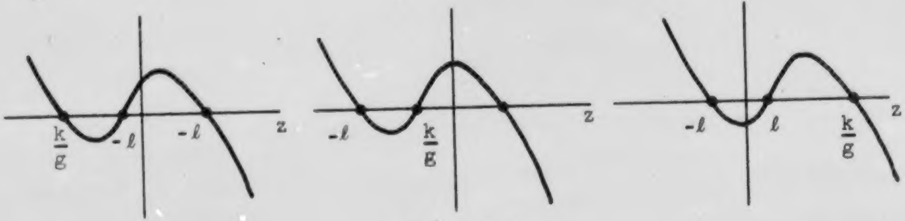


Figure 12-4c

Since the "total energy" $-\frac{\mu^2}{\ell^2}$ is negative and since the constraint implies $|z| \leq \ell$, the condition that the "potential energy" not exceed the "total energy" can be satisfied only if $\frac{k}{g} > -\ell$ and if $-\frac{\mu^2}{\ell^2}$ does not fall below the minimum of the "potential" in $[-\ell, \ell]$. If these conditions hold, the z-component of the motion is periodic and z ranges between a minimum value z_1 and a maximum value z_2. A half cycle of the oscillation is given by

$$t = \int_{z_1}^{z} \frac{\ell d\zeta}{\sqrt{2(k - g\zeta)(\ell^2 - \zeta^2) - \mu^2}} \tag{30}$$

and the period τ by

$$\tau = \int_{z_1}^{z_2} \frac{\ell\sqrt{2}\,d\zeta}{\sqrt{(k - g\zeta)(\ell^2 - \zeta^2) - \mu^2}} \,. \tag{31}$$

During one period θ increases by the amount

$$\Delta\theta = \int_0^{\tau} \theta' dt = 2 \int_{z_1}^{z_2} \frac{\theta'}{z'}\, dz$$

$$= 2\mu\ell \int_{z_1}^{z_2} \frac{d\zeta}{(\ell^2 - \zeta^2)\sqrt{2(k - g\zeta)(\ell^2 - \zeta^2) - \mu^2}} \,.$$

Unless $\Delta\theta$ is a rational multiple of π the motion of the pendulum will not be periodic. In Figure 12-4d we show how the orbit appears on surface of the sphere.

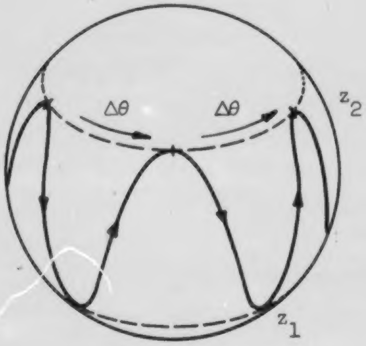


Figure 12-4d

The function $z \longrightarrow t$ given by (30) is called an elliptic integral and cannot be expressed as an elementary function. The inverse function $t \longrightarrow z$ is called an elliptic function. The elliptic integrals are so named because the integral for the arclength of an ellipse is a member of the class. These functions appear in many applications.

We have seen through a succession of problems of increasing complexity how it is possible to simplify the problem of specifying the motion of a system by use of constants of the motion (conservation laws) and constraints. With a sufficient number of such conditions we frequently can reduce the problem to one dimension. The techniques we have used for the solution of these problems will be found useful in many others.

Exercises 12-4

1. Show that if $\vec{K} = \vec{0}$ in (2) then the trajectory is a straight line. What information can you obtain from (7) in this case?

2. Use (7) or (8) to integrate the equations of motion for an inverse square force. (Hint: Replace ρ by $R = \frac{1}{\rho}$.)

3. Verify that the semi-minor axis b of the ellipse (12) is given by

$$b = \frac{mK^2}{\sqrt{\alpha^2 - \beta^2}} .$$

4. At what height will a satellite of the earth have the same period as the period of rotation of the earth about its axis? A synchronous satellite would be placed at this height. For the acceleration of gravity at the surface of the earth take $g = 980 \frac{\text{cm}}{\text{sec}^2}$ and for the radius of the earth, 6.37×10^8 cm.

5. What is the escape velocity from the earth's surface?

6. Consider a satellite in a circular orbit about the earth. Its retrorockets are fired briefly so that its speed is reduced but its direction of motion and position are changed negligibly. Suppose the change of speed is just enough to bring the satellite to the earth's surface. Without air resistance, what must the change in speed be and how long does it take the satellite to reach the earth after the retrorockets are fired?

7. Prove if $\mu = 0$ in (27) the motion of the spherical pendulum reduces to that of the circular pendulum of Section 12-3.

8. Under what conditions is the maximum value z_2 equal to the minimum value z_1 of z for the motion of a spherical pendulum? Discuss the motion in this case.

9. What is the motion of the pendulum when $z = 0$? (Hint: Equation (29) is not convenient for the study of this motion).

10. Determine the magnitude of the force of constraint $\lambda\vec{X}$ in (22).

Miscellaneous Exercises

1. Show that an object thrown with an upward component into a resisting atmosphere must come down to the same level at a speed less than its initial speed.

2. (a) Prove if the angular momentum of particles is conserved then the force is a central force.

   (b) Prove that a particle obeying Kepler's first two laws is subject to an inverse square force.

3. Show how the electric and magnetic field vectors $\vec{E}$ and $\vec{B}$ of Section 12-2(iii) must be modified to account for the motion of a charged particle in a coordinate frame which moves in translation with respect to the inertial frame with constant velocity $\vec{u}$.

4. The path of an object attracted by a central force is a circle and the center of force is on the circle. Show that the force law is an inverse fifth power law, $\vec{F} = \frac{\alpha \vec{X}}{\rho^6}$ and show that the speed is proportional to $\frac{1}{\rho^2}$.

5. (a) Let the path of a particle be given by $\vec{X} = \vec{r}(s)$ where s is arclength. Define the tangent $\vec{t}$ and principal normal $\vec{n}$ by $\vec{t} = \vec{r}'(s)$ and $\frac{d\vec{t}}{ds} = \kappa n$ where the sign of the curvature κ is nonnegative (as in Exercises 11-6, No. 19). Let $v = \frac{ds}{dt}$ and $\alpha = \frac{dv}{dt}$ be respectively, the speed and acceleration along the curve. Show that the force on the particle is $\vec{F} = m\alpha\vec{t} + m\kappa v^2 \vec{n}$.

   (b) Use the result of Part (a) to derive the inverse fifth power law in Number 4.

6. Consider a rocket which is to lift a payload of mass M_1 . Determine the amount of fuel relative to payload necessary to reach escape velocity in one minute from the earth given an exhaust velocity $v_c = 2 \times 10^5 \frac{\text{cm}}{\text{sec}}$ and a constant rate of fuel consumption. Neglect air resistance.

Chapter 13

NUMERICAL ANALYSIS

13-1. Introduction.

The creation of analysis was intimately associated with its applications and great mathematicians, Newton, Euler, Gauss, gave much of their thought to obtaining numerical solutions to applied problems. As analysis grew, mathematicians became increasingly concerned with general problems rather than specific problems and their numerical solution. As an example of a typical pattern for a modern theorem in analysis consider Theorem 6-3a* on the existence of the integral of a bounded function over an interval. The theorem refers to upper and lower estimates for the integral. If for every positive ϵ there exist upper and lower sums which lie within the tolerance ϵ of each other then the integral exists. This is the pattern: if the solution of a problem can be estimated then the solution exists. The problem is considered "solved" if it is shown that such estimation is "possible". To know that estimation is theoretically possible is not enough. In order to yield useable numerical results the method of estimation must be practicable by means at our command; furthermore, it should be economical and cost us no more in time and money than the solution is worth.

In recent years, with the advent of high speed electronic computation, the field of numerical analysis has interested mathematicians again. Their interest is not so much in specific computations as in the construction of effective general schemes of computation. Whether a given scheme is effective depends upon the instruments of computation at hand. Although such practical considerations have shaped the growth of numerical analysis there already is a body of elegant theorems in this area sufficient to capture the purest mathematical imagination.

In order to focus our thoughts, let us consider the problem of obtaining accurately to a given number of decimal places the period of a pendulum moving with the angular amplitude $\frac{\pi}{2}$. From Section 12-3, Equation (20b) we have for the period

---

*With slight modification, a result of Riemann, 1854.

141

(1) $$\tau = 2\sqrt{\frac{2\ell}{g}} \int_0^{\pi/2} \frac{d\psi}{\sqrt{\cos\psi}} = 2\sqrt{\frac{2\ell}{g}}\, I \,.$$

Suppose the factor $2\sqrt{\frac{2\ell}{g}}$ to be known to any desired accuracy and look only at the integral. Since the integrand is monotone we ought to be able to use upper and lower Riemann sums to estimate the integral. The integral is improper -- the integrand has an unbounded discontinuity at the upper end of integration -- but this is not a serious difficulty. We can obtain a continuous integrand by an appropriate substitution. For this purpose[*] we set $\cos\psi = (1-u)^2(1+2u) = g(u)$, $d\psi = \frac{6u(1-u)}{\sqrt{1-g(u)^2}}\,du$ and obtain

$$I = 6\int_0^1 \frac{u(1-u)}{\sqrt{[1-g(u)][1+g(u)]}\,\sqrt{g(u)}}\,du$$

$$= 6\int_0^1 \frac{du}{\sqrt{(3-2u)[1+g(u)](1+2u)}} \,.$$

The integrand is nowhere zero. Now the work of computation begins. We must find the intervals in which the integrand is monotone if we are to use the estimates by upper and lower sums as in Section 6-3(ii). In this case the integrand has a single extremum, a minimum, in the interior of the interval of integration. The zero of the derivative corresponding to this extremum is to be located, more usually approximated with a certain error. Now we can determine the minimum value of the function, again approximately, and fix the number of partition points needed in each of the two intervals of monotonicity to yield the desired accuracy. Next we may calculate the values of the integrand at each of the partition points, again to a degree of approximation. Finally the Riemann sums are calculated, perhaps with further errors introduced by rounding off to an appropriate number of decimal places. Thus even in this relatively straight forward example we see that a realistic problem may require a careful analysis of every arithmetical operation performed on the way to a numerical solution.

The problem is not always one of directly transforming a theoretical solution of a problem into a numerical solution as in the preceding example. Sometimes the theoretical solution cannot be translated into practical methods and new schemes must be devised. Furthermore, error analysis is not always practicable. Often computational schemes are used without any other justification except that they seem to work most of the time. Scientific and technological progress will not wait for sound error estimates. Empirical methods have taken us a long way. Still it is important to do what analysis we can. Naive empirical methods can be wasteful and time consuming; they are also known sometimes to give incorrect results unexpectedly.

---

[*] To get rid of the singularity at $\psi = \frac{\pi}{2}$, we insert the factor $(1-u)^2$. To avoid a singularity at the origin we make $g'(0) = 0$ by inserting the factor $1 + 2u$.

The literature of numerical analysis, or even that part of it accessible to us with tools of elementary calculus, is far too large for us to cover in the scope of a single chapter, or single volume, for that matter. Here we confine ourselves to a few basic methods: iteration schemes for finding the zeros of a function, approximation to a function in a neighborhood of a point, numerical integration, and numerical solution of differential equations.

Exercises 13-1

1. (a) Obtain a method for computing the square root of a positive number to within any prescribed tolerance ϵ. Obtain estimates to determine at what stage the process may be brought to an end.

   (b) Use the given method to obtain $\sqrt{7}$ accurate to 5 significant figures.

2. The idea of using Riemann sums to approximate the integral I given by Equation (1) has led us into formal complications. In such an approximation, the integrand is approximated by piecewise constant functions. Use a little more ingenuity in approximating the integrand and obtain upper and lower estimates for I. (Hint: note that $\int_x^{\pi/2} \frac{d\psi}{\sqrt{\cos \psi}} = \int_0^{\pi/2 - x} \frac{d\psi}{\sqrt{\sin \psi}}$ and use estimates for $\cos$ and $\sin$ obtained from Example 7-5(b).)

13-2. Iteration.

(i) Iterative methods for estimating a zero of a function.

An iterative scheme of approximation uses an estimate of a number to obtain a better one, the better estimate to obtain a still better one, and so on. For example, consider the problem of obtaining an approximation to $\sqrt{7}$ (Exercises 13-1, No. 1). We know that $a_0 = 3$ is an upper estimate for $\sqrt{7}$. It follows that $b_0 = \frac{7}{3}$ is a lower estimate. This suggests that we try the average of the two in order to improve the estimate:

$$a_1 = \frac{1}{2}(a_0 + b_0) = \frac{1}{2}(a_0 + \frac{7}{a_0}) .$$

With $a_0 = 3$ we have $a_1 = \frac{8}{3}$. Since $a_0^2 = 7 + 2$ and $a_1^2 = 7 + \frac{1}{9}$ we see that a_1 represents a considerable improvement over the estimate a_0. The same scheme can be repeated indefinitely: given an estimate a_k for $\sqrt{7}$ a better estimate a_{k+1} is obtained by the recursion formula

$$a_{k+1} = \frac{1}{2}(a_k + \frac{7}{a_k}) . \tag{1}$$

In particular, for $a_1 = \frac{8}{3}$, we obtain $a_2 = \frac{127}{48}$ with $a_2^2 = 7 + \frac{1}{2304}$. Let us see how good the improvement in the error is. If $a_k = \sqrt{7} + e_k$ then

$$\begin{aligned} e_{k+1} = a_{k+1} - \sqrt{7} &= \frac{1}{2}\left[\sqrt{7} + e_k + \frac{7}{\sqrt{7} + e_k}\right] - \sqrt{7} \\ &= \frac{1}{2}\left[\sqrt{7} + e_k + \frac{\sqrt{7}(\sqrt{7} + e_k) - e_k\sqrt{7}}{\sqrt{7} + e_k}\right] - \sqrt{7} \\ &= \frac{e_k^2}{2(\sqrt{7} + e_k)} . \end{aligned}$$

Since $\sqrt{7} > \frac{5}{2}$, and $e_k > 0$, we have

$$0 < e_{k+1} < \frac{e_k^2}{5} .$$

Consequently if a_k is accurate to n decimal places, that is, $e_k < \frac{10^{-n}}{2}$, then $e_{k+1} < \frac{10^{-2n}}{20} \leq \frac{10^{-2n-1}}{2}$ or a_{k+1} is accurate to more than twice the number of decimal places. Clearly, the scheme (1) is a very effective way of improving estimates for $\sqrt{7}$.

Iteration schemes are used to approximate a solution of an equation of the form

$$f(x) = 0 . \tag{2}$$

Thus, the equation $x^2 - 7 = 0$ can be solved approximately within any given tolerance by (1). The method is to recast (2) in the form

$$x = \phi(x) , \tag{3}$$

and to use the iteration scheme

$$a_{k+1} = \phi(a_k) , \tag{4}$$

to approximate the desired solution of (2) and (3). The estimate a_k is called the k-th iterant of the scheme. For the example $f : x \to x^2 - 7$ we have used $\phi : x \to \frac{1}{2}(x + \frac{7}{x})$. There are many possible ways to obtain an equation of the Form (3) which is equivalent to (2). The problem is to obtain an iteration scheme which works, that is, one for which the approximation is improved at each step. Moreover we want the improvement to be substantial so that it is worth the effort of computation.

If a is a solution of (2), that is $f(a) = 0$, then a is also a solution of

$$x = \phi(x) = x + g(x)f(x) \tag{5}$$

where g is any function defined at a . We seek a choice of the function g so that the iteration scheme

$$a_{k+1} = a_k + g(a_k)f(a_k) \tag{6}$$

converges to a , namely, so that

$$\lim_{k \sim \infty} \phi(a_k) = a . \tag{7}$$

Furthermore, since the only kind of condition we can usually impose on the initial estimate a_0 is that it be within some neighborhood of a , we should like (7) to hold, independently of the choice of initial estimate, within a neighborhood of a .

Now, let us suppose

$$a = \phi(a)$$

and ask under what circumstances the estimate a_{k+1} given by (4) is a better estimate of a than a_k . Set $e_k = a_k - a$. The criterion for improvement in the estimate is $\left|\frac{e_{k+1}}{e_k}\right| < 1$, or

(8) $$\left|\frac{e_{k+1}}{e_k}\right| = \left|\frac{a_{k+1} - a}{a_k - a}\right| = \left|\frac{\phi(a_k) - \phi(a)}{a_k - a}\right| < 1 .$$

Geometrically this criterion is that the steepness (or absolute slope) of the chord to the graph $y = \phi(x)$ between the points $(a, \phi(a))$ and $(a_k, \phi(a_k))$ is less than 1 . This means that $(a_k, \phi(a_k))$ lies in either the right or left quadrant bounded by the lines of slope ± 1 through the point $(a, \phi(a)) = (a, a)$, the unshaded region in Figure 13-2a.

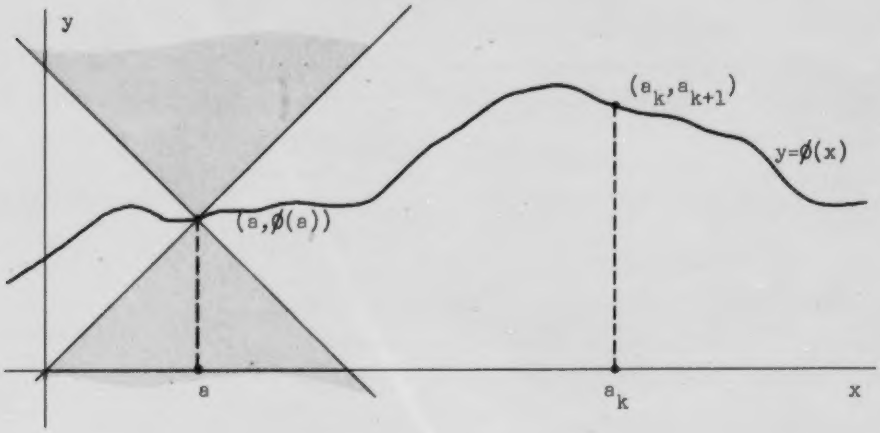


Figure 13-2a

If a_k lies within a neighborhood of a such that the graph $y = \phi(x)$ lies within this region for all points of the neighborhood then (8) is satisfied not only by a_k , but also by all succeeding iterants. If ϕ is differentiable we observe that the graph lies within the region for all x in some neighborhood I of a if, and only if, $\left|\frac{\phi(x) - \phi(a)}{x - a}\right| < 1$ for all x in I . For this it is sufficient that $|\phi'(x)| < 1$ on I , an immediate consequence of the Law of the Mean. Finally, note if the steepness of the graph is bounded below 1 on I ,

(9) $$|\phi'(x)| \leq M < 1 ,$$

then, for $a_0 \in I$,

$$\begin{aligned}
|e_0| &= |a_0 - a| , \\
|e_1| &= |a_1 - a| = |\phi(a_0) - \phi(a)| \\
&= |\phi'(\xi_0)(a_0 - a)| \\
&\leq M|e_0| , \\
|e_2| &= |a_2 - a| = |\phi(a_1) - \phi(a)| \\
&= |\phi'(\xi_1)(a_1 - a)| \leq M|e_1| \\
&\leq M^2|e_0| ,
\end{aligned}$$

and, in general,

$$|e_k| \leq M^k|e_0| . \tag{10}$$

Thus the iteration scheme converges and the error bound approaches zero geometrically.

With (10) in mind it is clearly in our interest to choose a function g in (5) to make M as small as we conveniently can. First let us see what can be done with a constant function $g : x \to c$. We can require that $\phi'(a) = 0$, and if f is continuously differentiable this will mean that M can be held within any desired tolerance by restriction to a sufficiently small neighborhood of a . From $\phi(x) = x + cf(x)$ this yields $\phi'(a) = 1 + cf'(a)$, or $c = -\frac{1}{f'(a)}$. Since we do not know a exactly (otherwise we would not be looking for an approximation to a), we generally do not know $f'(a)$ exactly, and in fact any reasonable approximation to $-\frac{1}{f'(a)}$ may do for c . Thus for the solution of (2) we are led to the iteration scheme

$$a_{k+1} = a_k + cf(a_k) , \qquad \left(c \approx -\frac{1}{f'(a)}\right) . \tag{11}$$

Example 13-2a. Let us calculate the real zero of $f : x \to x^3 + x - 1$. Since $f'(x) = 3x^2 + 1 > 0$, the function f is increasing and the zero is unique. Since $f(0) = -1$ and $f(1) = 1$ we take $a_0 = \frac{1}{2}$ and approximate $f'(a)$ in (11) by $f'(\frac{1}{2}) = \frac{7}{4}$. Thus we employ the iteration scheme

$$\begin{aligned}
a_{k+1} &= a_k - \tfrac{4}{7}(a_k{}^3 + a_k - 1) \\
&= \tfrac{1}{7}(3a_k - 4a_k{}^3 + 4)
\end{aligned}$$

to obtain with $a_0 = 0.5$, $f(0.5) = -\frac{3}{8} = -.375$

$$a_1 = \frac{5}{7} = 0.7 \ldots , \quad f(0.7) = .043 ,$$

$$a_2 = \frac{1607}{2401} = 0.67 \ldots , \quad f(0.67) = -.03 ,$$

and so on.

Sometimes a seemingly slight modification will result in a great improvement in an iteration scheme. Sometimes an algebraic device will reduce the amount of computation necessary. The following example illustrates both.

Example 13-2b. Consider the equation

$$1 + x + x^2 + x^3 + \ldots + x^{10} = 12.$$

There is a solution between $x = 1$ and $x = 2$ and we may expect it to be much closer to 1 . We sum the geometrical progression to recast the equation in the form

$$\frac{x^{11} - 1}{x - 1} = 12 \tag{12a}$$

or

$$f(x) = x^{11} - 12x + 11 = 0 . \tag{12b}$$

Now if a is the desired solution, $f(a) = 0$, we have

$$\begin{aligned} f'(a) &= 11a^{10} - 12 \\ &= \frac{11a^{11}}{a} - 12 \\ &= \frac{11}{a}(12a - 11) - 12 \end{aligned}$$

or

$$f'(a) = 11(12 - \frac{11}{a}) - 12 . \tag{13}$$

Since a is close to 1 we may suppose that $f'(a) \approx f'(1) = -1$. With $f'(a)$ replaced by -1 in (11) this yields the iteration scheme

$$\begin{aligned} a_{k+1} &= a_k + (a_k^{11} - 12a_k + 11) \\ &= a_k^{11} - 11a_k + 11 . \end{aligned}$$

Now if we use $a_0 = 1$ as our first approximation to a we obtain $a_1 = 1$. The successive iterants are all equal to 1 and the scheme is not useful for solving the original problem. Where have we been careless?

In multiplying by $x - 1$ to obtain (12b) we introduced the extraneous solution $x = 1$ which is close to the solution we are seeking. Moreover, $f'(x) = 0$ at $x = \sqrt[10]{\frac{12}{11}}$ which also is close to 1 ; so the approximation to $\frac{1}{f'(a)}$ may be far from the truth. If we are to benefit from our clever algebra and use (12) we shall evidently have to begin with a better initial estimate for a.

To obtain an initial estimate for a, set $x = 1 + u$ in (12a):

$$\frac{(1 + u)^{11} - 1}{u} = 12.$$

Now use the binomial theorem and divide by u to obtain

$$11 + 55u + 165u^2 + \ldots = 12 .$$

Since we expect u to be close to zero we expect higher powers of u to be small compared with u and we ignore powers higher than the first to obtain for an initial approximation

$$11 + 55u_0 = 12 .$$

Thus $1 + u_0 = \frac{56}{55} = 1.018 \ldots$ and we take $a_0 = 1.02$. Now we insert this value instead of a in (13) to obtain $f'(a) \approx 1.4 \approx \frac{10}{7}$ (Note that there is no point to carrying many decimal places initially). With this estimate replacing $f'(a)$ in (11) we obtain the iteration scheme

$$a_{k+1} = a_k - \frac{7}{10}(a_k^{11} - 12a_k + 11) .$$

To calculate a_k^{11} we use a log table or the LL scale on a slide rule and obtain successively the iterants $a_0 = 1.02$, $a_1 = 1.0175$, $a_2 = 1.0172$. In two iterations we reach the limits of slide rule accuracy.

In this case we have seen how a proper initial estimate may be crucial for the convergence of an iteration scheme. For many cases to obtain a suitable first estimate is half the battle.

The most commonly used scheme for which the function g of (5) is not constant is Newton's Method. In Newton's Method $g(x) = -\frac{1}{f'(x)}$; thus

$c \approx -\frac{1}{f'(a)}$ in (11) is replaced by $-\frac{1}{f'(a_k)}$ to give the scheme

$$a_{k+1} = a_k - \frac{f(a_k)}{f'(a_k)} \,. \tag{14}$$

The iteration scheme used for $\sqrt{7}$ above is Newton's Method (14) with $f : x \to x^2 - 7$.

One major advantage of a convergent interation scheme is that it is self-correcting. An error in computation at any stage, provided it is not too large, will not prevent convergence to the solution. The computational methods of elementary arithmetic do not have this property. Because of this property slight modifications in the scheme, either systematic ones or special adaptations made at any step in the course of the computation will not affect the outcome. For example, the derivative $f'(a_k)$ in Newton's Method (14) may be harder to compute than the function values. In that case it may be more efficient in terms of the effort of computation to replace the derivative by a difference quotient as in

$$a_{k+1} = a_k - f(a_k) \Big/ \frac{f(a_k) - f(a_{k-1})}{a_k - a_{k-1}} \,. \tag{15}$$

Any combination of the methods we have described may also be adopted. For example, it may be convenient to use (14) or (15) until we are close to the desired level of accuracy say up to $k = m$ and then use (11) for the final stages of the computation, $k > m$, with the constant c taken as $-\frac{1}{f'(a_m)}$ or $\frac{a_m - a_{m-1}}{f(a_m) - f(a_{m-1})}$, whichever is appropriate.

(ii) Picard's method for the solution of a differential equation.

The idea of iteration is just as important for theoretical questions as it is for numerical computation. If an iteration scheme converges for a given equation to a value a in the domain of a continuous function ϕ , this fact alone shows that a solution of (3) exists; namely $\phi(a) = a$ (Exercises 13-2, No. 6). Furthermore, iterative methods can be applied in situations of other kinds. Consider, for example, the following iterative solutions, Picard's Method*, for the differential equation

---

*E. Picard, French, (1893). The method was first published by Liouville in 1838. Picard greatly generalized the method.

(16a) $$\frac{dy}{dx} = \Phi(x,y)$$

subject to the initial condition

(16b) $$y = y_0 \quad \text{at} \quad x = x_0 \; .$$

By a solution, we mean a function $u : x \to y$ for which $u'(x) = \Phi(x, u(x))$ and $u(x_0) = y_0$. For the solution u we then have

(17) $$u(x) = y_0 + \int_{x_0}^{x} \Phi(x, u(x))dx \; .$$

This form of (16) suggests the iteration

(18a) $$u_0(x) = y_0 \; ,$$

(18b) $$u_{k+1}(x) = y_0 + \int_{x_0}^{x} \Phi(x, u_k(x))dx \; .$$

It is easy to give general sufficient conditions on the function Φ for the convergence of this scheme (Exercises 14-M, No. 15) and to show under these conditions that the successive iterants converge to a solution of (12) and that the solution is unique. For this we need some of the ideas of Chapter 14. Here we use the scheme to prove uniqueness for the solution under the condition that the derivative be bounded stated in Section 10-9 (p. 625).

THEOREM 13-2. Let Equation (16a) be separable, $\Phi(x,y) = f(x)g(y)$. If f is continuous on a neighborhood of x_0 and g' bounded on a neighborhood of y_0 , then there is exactly one solution $u : x \to y$ of (16a) satisfying the initial condition (16b) in a neighborhood of x_0 .

Proof. Observe that we have already demonstrated the existence of a solution (Theorem 10-9), the only question is that of uniqueness. Let u_k be an approximation to a solution u on a neighborhood of x_0 where for the error $\alpha_k(x)$, we have

(19) $$|\alpha_k(x)| = |u_k(x) - u(x)| < \epsilon_k$$

now

$$\begin{aligned}\alpha_{k+1}(x) &= u_{k+1}(x) - u(x)\\ &= \int_{x_0}^{x} f(\xi)[gu_k(\xi) - gu(\xi)]d\xi\\ &= \int_{x_0}^{x} f(\xi)g'(\eta)[u_k(\xi) - u(\xi)]d\xi\end{aligned}$$

where η is a number between $u_k(\xi)$ and $u(\xi)$ according to the Law of the Mean. Now since u_0 is continuous and satisfies the initial condition (16b), then by (18) u_k is continuous (since it is differentiable) and also satisfies (16b) for $k \geq 1$. From the continuity of u and u_k there is a neighborhood of x_0 where the values of both functions lie within the neighborhood of y_0 where $g'(y)$ is bounded; hence $g'(\eta)$ has the same bound. Furthermore, the neighborhood of x_0 can be taken within a closed interval where $f(x)$ is bounded. Consequently, on some neighborhood of x_0 we may satisfy $|f(x)g'(\eta)| < M$. We may also restrict this neighborhood so that (19) is satisfied and so obtain

(20) $$|\alpha_{k+1}(x)| \leq \epsilon_{k+1} = |x - x_0|M\epsilon_k .$$

In order to guarantee a reduction in the error we need only require $|x - x_0|M < r < 1$; then

(21) $$|\alpha_k(x)| \leq \epsilon_k < r\epsilon_{k-1} < r^2\epsilon_{k-2} < \ldots < r^k\epsilon_0$$

where ϵ_0 is the error tolerance for the initial approximation (the details of verification are left to Exercises 13-2, No. 7). Thus on some interval about x_0, the iteration scheme converges for each value of x to $u(x)$. It follows that there can only be one solution u to (16) under the hypotheses of the theorem as we now prove. Let I be a neighborhood of x_0 where u and v are solutions of (16) and where the hypotheses of the theorem hold for u (in particular, that $g'u(x)$ is bounded). Since both u and v are solutions we have proved that there is a neighborhood of x_0 where the iteration scheme (18) converges to $u(x)$ and a neighborhood where it converges to $v(x)$; hence on the smaller of these neighborhoods, say J, the iteration scheme converges to both $u(x)$ and $v(x)$; hence, on J, $u(x) = v(x)$. Now, if $u(x) \neq v(x)$ on I there must be at least one point ξ in I such that $u(\xi) \neq v(\xi)$. If some such point exists with $\xi > x_0$, then consider

(22) $$x_1 = \text{glb}[\xi : u(\xi) \neq v(\xi) \text{ and } \xi > x_0] .$$

We then have $x_1 > x_0$ since we have found a neighborhood J of x_0 where $u(x) = v(x)$. Furthermore, $x_1 \in I$ since I contains at least one point $\xi > x_0$ where $u(\xi) \neq v(\xi)$, and since x_1 is the greatest lower bound of the set of such points. Because u and v are continuous and $u(x) = v(x)$ for $x < x_1$, it follows that the two are equal at $x = x_1$, say $y_1 = u(x_1) = v(x_1)$. Further, since $x_1 \in I$, the hypothesis of the theorem holds there. Now, u and v are both solutions of (16a) satisfying the initial condition $y = y_1$ at $x = x_1$; hence as we have argued above $u = v$ on some neighborhood of x_1. Hence x_1 cannot be the greatest lower bound defined by (22). We conclude that $u(x) = v(x)$ for $x > x_0$ in I and by a similar argument that $u(x) = v(x)$ for $x < x_0$ on I. Thus u and v are the same.

Exercises 13-2

1. Devise an iteration scheme to approximate $\sqrt{A}$ for any positive A. Show how to choose an initial estimate for $\sqrt{A}$ so that the scheme converges and verify that the error can be brought below any given tolerance.

2. (a) An iteration scheme is called alternating if the error changes sign at each iteration. Since any two consecutive iterants approximate the solution from above and below, the value of an alternating scheme is that it permits an estimate of the error without a separate error analysis. Find a sufficient condition that a convergent iteration scheme be alternating.

   (b) Construct a convergent alternating scheme for calculating $\sqrt{A}$.

   (c) Use the alternating scheme obtained in Part (b) to calculate $\sqrt{3}$ accurately to two decimal places.

3. Obtain an iteration scheme for $\sqrt[n]{A}$, demonstrate convergence, and estimate the error in the k-th iterant.

4. (a) Prove the following theorem. Suppose $f(a) = 0$. If f has two continuous derivatives on a neighborhood of a and if $f'(a) \neq 0$, then Newton's Method (14) converges if the initial estimate is sufficiently close to a. (Hint: use the Law of the Mean twice to approximate $f(a_k)$ as for the tangent approximation, Section 5-7).

(b) Show that Newton's Method is monotone (the error has constant sign) if in addition to the conditions of Part (a), $f''(a) \neq 0$.

5. Obtain the greatest zero of f to 3 decimal places, where

(a) $f(x) = x^6 + 6x - 9$;

(b) $f(x) = x^3 - 4x^2 + 4x - 2$;

(c) $f(x) = \sum_{n=0}^{15} x^n = 48$.

6. Show if the iteration scheme (4) converges to a number a in the domain of ϕ , and if a is a point of continuity of ϕ , that a is a solution of (3); that is, $\phi(a) = a$.

7. Verify the error estimate (17) in the Picard iteration scheme for a separable equation under the conditions of Theorem 13-2.

8. Consider the differential equation $y' = 4x\sqrt{y}$ which has more than one solution $u : x \to y$ satisfying the initial condition $u(x) = 0$ at $x = 0$, for example, $u : x \to x^4$ and $u : x \to 0$. In view of Theorem 13-2, how can this be possible?

13-3. Taylor's[*] Theorem with Remainder.

In Section 5-7 we considered the tangent approximation to a function f in the neighborhood of a point a ; namely, as an approximation to f we took the linear function

$$f_1 : x \to f(a) + f'(a)(x - a)$$

which has the same value and slope as f at a . We may reasonably suppose that a function which has the same derivatives as f up to n-th order, where $n > 1$, is a better approximation to f in some neighborhood of a . (It is convenient in this context to call $f(a)$ the zero-order derivative of f and we shall frequently write $f^{(0)}(a)$ for $f(a)$). We consider such an approximation in the form of a polynomial function

$$\text{(1a)} \qquad f_n : x \to c_0 + c_1(x - a) + c_2(x - a)^2 + \dots + c_n(x - a)^n .$$

The $n + 1$ conditions, $f_n^{(k)}(a) = f_n(a)$, for $k = 0 , 1 , 2 , \dots , n$, determine the $n + 1$ coefficients c_k : from

$$f_n^{(k)}(x) = k!c_k + \frac{(k + 1)!}{1!} c_{k+1}(x - a) + \dots + \frac{n!}{(n - k)!} c_n(x - a)^{n-k}$$

we have, on setting $x = a$,

$$\text{(1b)} \qquad c_k = \frac{f^{(k)}(a)}{k!} , \qquad (k = 0,1,\dots,n)$$

Entering these expressions for the coefficients in (1a), we obtain

$$\text{(2)} \quad f_n(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2 + \dots + \frac{f^{(n)}(a)}{n!}(x - a)^n$$

$$= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x - a)^k ,$$

the so-called Taylor polynomial of n-th order at a

In Section 5-7 we obtained an estimate of the error of the tangent approximation in terms of a bound on the second derivative. For the approximation (2) we obtain an error estimate in terms of a bound on the $(n + 1)$-th derivative.

---

[*] Brook Taylor. English 1685-1731.

THEOREM 13-3. (Taylor's Theorem) Let f be a function with $n + 1$ continuous derivatives on the closed interval I with endpoints a and b. If $|f^{(n+1)}(x)| \leq M_{n+1}$ on I, then

$$f(b) = f(a) + f'(a)(b - a) + \dots + \frac{f^n(a)}{n!}(b - a)^n + R_n(b) , \tag{3}$$

where the remainder $R_n(b)$ is bounded by

$$|R_n(b)| \leq \frac{M_{n+1}}{(n + 1)!}|b - a|^{n+1} . \tag{4}$$

Proof. For the proof we use the method already employed in Example 7-5b to approximate sin and cos and in Section 8-6 to approximate the exponential function. Suppose $b > a$, (the proof for $b < a$ is left to Exercises 13-3, No. 5). We have for $x \in [a,b]$,

$$-M_{n+1} < f^{(n+1)}(x) < M_{n+1} . \tag{5}$$

Integrate from a to t in (5), where $t \in [a,b]$, and apply Theorem 6-4a to obtain

$$-M_{n+1}(t - a) \leq f^{(n)}(t) - f^{(n)}(a) \leq M_{n+1}(t - a) ;$$

whence

$$f^{(n)}(a) - M_{n+1}(t - a) \leq f^{(n)}(t) \leq f^{(n)}(a) + M_{n+1}(t - a) . \tag{6}$$

Now integrate again from a to x, where $x \in [a,b]$, and obtain similarly

$$\begin{aligned} f^{(n-1)}(a) + f^{(n)}(a)(x - a) - M_{n+1}\frac{(x - a)^2}{2} &\leq f^{(n-1)}(x) \\ &\leq f^{(n-1)}(a) + f^{(n)}(a)(x - a) + M_{n+1}\frac{(x - a)^2}{2} . \end{aligned} \tag{7}$$

Note that this is just the tangent approximation for $f^{(n-1)}$. If we set $n = 1$ in (7) we obtain a sharper estimate for the tangent approximation then that of (1) in Section 5-7, namely

$$|R_1(x)| \leq \frac{1}{2} M_2(x - a)^2 . \tag{8}$$

We may integrate repeatedly in the same way to obtain higher order approximations from (5) and to obtain the result of the theorem. The completion of the proof by mathematical induction is left to Exercises 13-3, Number 5.

Example 13-2a. We calculate $\sqrt{1.01}$ accurately to six decimal places. For this use the function $f : x \to \sqrt{1 + x}$ with $x = \frac{1}{100}$. We have for the successive derivatives

$$f'(x) = \frac{1}{2\sqrt{1 + x}}$$

$$f''(x) = - \frac{1}{4(1 + x)^{3/2}}$$

$$f'''(x) = \frac{3}{8(1 + x)^{5/2}} .$$

Observe, since $f'''(x)$ is decreasing that for $x \in [0, \frac{1}{100}]$,

$$|f'''(x)| \le |f(0)| \le \frac{3}{8} .$$

Consequently, $|R_2(\frac{1}{100})| \le \frac{3}{8}(\frac{1}{100})^3 \frac{1}{3!} \le \frac{10^{-6}}{16}$. Thus the second order approximation $f_2(\frac{1}{100})$ yields the desired accuracy. We obtain

$$\begin{aligned} \sqrt{1.01} &\approx 1 + \frac{1}{2}(\frac{1}{100}) - \frac{1}{4}(\frac{1}{10,000})(\frac{1}{2!}) \\ &\approx 1.0049875 . \end{aligned}$$

Round-off in the wrong direction would create an error of more than half a unit in the sixth decimal place. At this point we do not know which way to round off because we do not know the sign of the error. However, observe that $f'''(\frac{1}{100}) > 0$. We anticipate that $f_3(\frac{1}{100})$ is a better approximation than $f_2(\frac{1}{100})$, therefore we round upward to obtain

$$\sqrt{1.01} \approx 1.004988 .$$

The proof that this is correct is left to Exercises 13-3, Number 6.

Example 13-3b. Next we show how to approximate arcsin x for small positive values of x . It would be possible to calculate the successive derivatives of arcsin x and use the theorem directly. We shall proceed slightly differently and use

$$\arcsin x = \int_0^x \frac{dt}{\sqrt{1 - t^2}} .$$

First, we apply Taylor's Theorem to

$$g(u) = \frac{1}{\sqrt{1 - u}}$$

at $u = 0$. For the derivatives of u we have successively

$$g'(u) = \frac{1}{2(1 - u)^{3/2}}$$

$$g''(u) = \frac{3}{4(1 - u)^{5/2}}$$

$$- - -$$

$$g^{(k)}(u) = \frac{1 \cdot 3 \cdot 5 \cdots (2k - 1)}{2^k(1 - u)^{(2k+1)/2}} = \frac{(2k)!}{k!2^{2k}} \frac{1}{(1 - u)^{(2k+1)/2}} .$$

Since $g^{(k)}$ is increasing we see that the maximum on $[0,u]$ is reached at u. It follows that

$$g(u) = g_n(u) + Q_n(u) = \sum_{k=0}^{n} \frac{(2k)!u^k}{(k!)^2 2^{2k}} + Q_n(u)$$

where

$$|Q_n(u)| \leq \frac{1 \cdot 3 \cdot 5 \cdots (2n - 1)(2n + 1)}{(n + 1)! \, 2^{n+1}} \frac{u^{n+1}}{(1 - u)^{(2n+3)/2}}$$

$$\leq \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdots \frac{2n + 1}{2n + 2} \frac{u^{n+1}}{(1 - u)^{n+1}}$$

$$\leq \frac{1}{2}\left(\frac{u}{1 - u}\right)^{n+1} \frac{1}{\sqrt{1 - u}} .$$

Now restrict u so that

$$\frac{u}{1 - u} = r < 1 . \tag{9}$$

This will occur only if $u < \frac{1}{2}$, hence $\sqrt{1 - u} > \frac{1}{\sqrt{2}}$. Under this condition, then,

$$|Q_n(u)| < \frac{\sqrt{2}}{2} r^{n+1} < r^{n+1} , \tag{10}$$

and the remainder can be brought below any tolerance for sufficiently large n. Subject to (9) we then have, with $u = t^2$,

$$\frac{1}{\sqrt{1 - t^2}} = g(t^2) = \sum_{k=0}^{n} \frac{(2k)! \, t^{2k}}{(k!)^2 \, 2^{2k}} + Q_n(t^2) .$$

Since $\frac{1}{\sqrt{1 - t^2}}$ and the polynomial $g_n(t^2)$ are both integrable it follows by Theorem 6-4c that $Q_n(t^2)$ is integrable. We integrate from 0 to x, where $|x| < \frac{1}{\sqrt{2}}$, to obtain

$$\arcsin x = \sum_{k=0}^{n} \frac{(2k)!\ t^{2k+1}}{(2k + 1)(k!)^2 2^{2k}} + R_n(x) ,$$

where, from (9) and (10),

$$(11) \quad |R_n(x)| = \left|\int_0^x Q_n(t^2)dt\right| < \left|\int_0^x \left(\frac{t^2}{1 - t^2}\right)^{n+1} dt\right| < |x|\left(\frac{x^2}{1 - x^2}\right)^{n+1} < 2^{n+1}x^{2n+2} ,$$

where we have used $\frac{x^2}{1 - x^2}$ as an upper bound for $\frac{t^2}{1 - t^2}$ on $[0,x]$. Sharper error estimates are possible and $|R_n(x)|$ can be made less than any tolerance for $|x| < 1$, not just $|x| < \frac{1}{2}$, by taking n large enough (see Exercises 13-3, No. 9).

The preceding example offers an instance of the representation of a function in a neighborhood of a given point as the limit of its Taylor polynomials at the point. If we can make the remainder $R_n(x)$ smaller than any given tolerance for each point x in some neighborhood of a simply by taking n large enough, then

$$(12) \qquad f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x - a)^k .$$

In that case we write

$$(13) \quad f(x) = \sum_{n=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x - a)^k$$

$$= f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x - a)^n + \ldots ,$$

where by the infinite sum in (13) we mean the limit of the finite sum in (12). The sum in (13) is called the Taylor series or Taylor expansion of f in the neighborhood of a. Taylor did not obtain an estimate of the remainder as in Theorem 13-3, but described the technique of expansion. Expressions for the remainder were first given by Lagrange and Cauchy (see Exercises 13-3, Nos. 8 and 9). We shall postpone the general question of existence of a Taylor expansion to Chapter 14 where we develop better methods of handling the issue.

Exercises 13-3

1. How many terms of the Taylor expansion of $\sqrt{1 - x}$ in the neighborhood of $x = 0$ should be used to give $\sqrt{7} = \frac{8}{3}\sqrt{1 - \frac{1}{64}}$ accurately to five decimal places?

2. Obtain $\sqrt[3]{9}$ to three decimal place accuracy.

3. Give $\tan \frac{1}{100}$ accurately to 5 decimal places.

4. Give the third order Taylor polynomial at $x = a$ in each following case. Obtain a formula for the general term if you perceive the pattern.

   (a) $\sqrt{1 + x}$, $a = 0$

   (b) $\sqrt{1 + x^2}$, $a = 0$

   (c) $\tan x$, $a = 0$

   (d) $\frac{1}{\cos x}$, $a = 0$

   (e) $\frac{1}{\sin x}$, $a = \frac{\pi}{2}$

   (f) $\log x$, $a = 1$

   (g) $\frac{1}{\sqrt{1 + x}}$, $a = 0$

   (h) $\sinh \log x$, $a = 1$

   (i) $4x^3 + 7x^2 + 2x + 5$, $a = -1$

   (j) $\log \cos x$, $a = 0$

5. (a) Complete the proof of Theorem 13-3 by induction for $b > a$. The case $b = a$ is trivial.

   (b) Give the proof for the case $b < a$.

6. Show if $f^{(n+1)}(x)$ has constant sign in the interval I of Theorem 13-3 then the remainder $R_n(x)$ has the same sign as

$$f^{n+1}(x)(b - a)^{n+1} .$$

7. In Example 13-3b we found approximating polynomials for arc sin. We did not actually prove that these are Taylor polynomials by verifying that the coefficients satisfy (1b). Show that these are Taylor polynomials by proving the following general uniqueness theorem. If f has $n + 1$ continuous derivatives and there exists a neighborhood of a where

$$f(x) = c_0 + c_1(x - a) + \ldots + c_n(x - a)^n + Q_n(x)$$

where $|Q_n(x)| \leq K_1|x - a|^{n+1}$, then

$$c_k = \frac{f^{(k)}(a)}{k!} , \qquad (i = 0,1,2,\ldots,n) .$$

8. Lagrange obtained a form of the remainder $R_n(b)$ in (3) which generalizes the Law of the Mean. For this, apply the Law of the Mean to the function

$$\text{(i)} \qquad g(x) = \sum_{k=0}^{n} \frac{(b - x)^k}{k!} f^{(k)}(x) + A(b - x)^{n+1}$$

on the interval $[a,b]$ with the constant A chosen so that $g(a) = g(b) = f(b)$ and verify that $A = \frac{f^{(n+1)}(\xi)}{(n + 1)!}$ where ξ lies between a and b. Thus the Lagrange form of the remainder is $R_n(x) = \frac{f^{(n+1)}(\xi)}{(n + 1)!}(x - a)^{n+1}$. What conditions must f and its derivatives satisfy for this result?

∧9. (a) Show that the remainder in Taylor's Theorem can be written in the form

$$R_n(b) = \frac{1}{n!} \int_a^b (b - x)^n f^{n+1}(x)dx ,$$

where it is assumed that f has $n + 1$ continuous derivatives. (Hint: use induction and integration by parts. Compare Chapter 10, Miscellaneous Exercises, No. 20.)

(b) From the integral form of the remainder obtain Cauchy's remainder

$$R_n(b) = \frac{1}{n!}(b - a)(b - u)^n f^{(n+1)}(u)$$

where u lies between a and b.

(c) Use Cauchy's form of the remainder to prove the claim of the text for the Taylor expansion of arcsin in Example 13-3, that $\lim_{n\sim\infty} R_n(x) = 0$ for $|x| < 1$.

10. The eccentricity e of an ellipse is given by $e^2 = 1 - \frac{b^2}{a^2}$ where a is the semi-major axis and b, the semi-minor axis. The circle, $a = b$, has eccentricity $e = 0$, thus e measures the departure from circular symmetry. Obtain the arclength of the ellipse in the form $s = a\, f(e)$ and expand f in powers of e to sixth order. (As we mentioned in Section 12-4(iii), the integral for the arclength of an ellipse cannot be written in terms of elementary functions. The solution of this exercise yields precise estimates of the arclength provided the eccentricity is not too large.)

11. A function f is said to have a zero of order k at $x = a$ if $0 = f(a) = f'(a) = f''(a) = \dots = f^{(k-1)}(a)$ and $f^{(k)}(a) \neq 0$; the leading term in the Taylor expansion of f at a is then $\frac{f^{(k)}(a)}{n!}(x - a)^k$. Prove if f has a first order zero at $x = a$ then the function g given by $g(x) = [f(x)]^n$ has a zero of order n . (Hint: Use the Lagrange remainder of No. 8 for f .)

12. Two curves are said to have a contact of order n at a point X_0 if n is the largest integer for which the curves have parametrizations $\vec{X} = \vec{r}(t)$, $\vec{Y} = \vec{q}(t)$, respectively, with $\vec{X}_0 = \vec{r}(t_0) = \vec{q}(t_0)$ such that $\vec{r}'(t_0) \neq \vec{0}$ and $\vec{q}'(t_0) \neq \vec{0}$ and $\vec{r}^{(k)}(t_0) = \vec{q}^{(k)}(t_0)$ for $k = 0, 1, \dots, n$ and $\vec{r}^{(n+1)}(t_0) \neq \vec{q}^{(n+1)}(t_0)$. Taylor's Theorem can easily be extended component-by-component to vector functions so that this condition may also be given in terms of Taylor polynomials as before.

(a) Prove that if t is replaced by an equivalent parameter, the order of contact is unaffected.

(b) Let s and σ be arclength along the curves $\vec{X} = \vec{r}(t)$ and $\vec{Y} = \vec{q}(t)$. Show if the curves have contact of order n as defined in Part (a) then the parameter of Part (a) may be replaced by arclength; i.e., for

$$s = \int_{t_0}^{t} |r'(\tau)|\,d\tau \quad \text{and} \quad \sigma = \int_{t_0}^{t} |q'(\tau)|\,d\tau ,$$

we have $\left.\frac{d^k\vec{X}}{ds^k}\right|_{s=0} = \left.\frac{d^k\vec{Y}}{d\sigma^k}\right|_{\sigma=0}$ where $k = 0, 1, 2, \dots, n$,

and

$$\left.\frac{d^{n+1}\vec{X}}{ds^{n+1}}\right|_{s=0} \neq \left.\frac{d^{n+1}\vec{Y}}{d\sigma^{n+1}}\right|_{\sigma=0}$$

(c) Show that the curves $y = f(x)$ and $y = g(x)$ have a contact of order n at $x = a$ if, and only if, $f - g$ has a zero of order $n + 1$ at a .

13. Infinite order of contact at a given point does not necessarily imply that the two curves coincide on any neighborhood of the point. Show that the curve

$$y = \begin{cases} e^{-1/x^2} & \text{, for } x \neq 0 \\ 0 & \text{, for } x = 0 \end{cases}$$

has a contact of infinite order with the x-axis at $x = 0$.

14. (a) Show that the osculating circle to a curve at a given point has a contact of order 2 or more.

(b) Prove that if an osculating circle to a plane curve has a contact of order 2 then it crosses the curve at the point of contact.

15. Given $y = f(x)$ and $y = g(x)$ have contact of order n at $x = a$ and $g^{(n+1)}(a) \neq 0$, prove that

$$\lim_{x\sim a} \frac{f(x)}{g(x)} = \lim_{x\sim a} \frac{f'(x)}{g'(x)} = \lim_{x\sim a} \frac{f''(x)}{g''(x)} = \dots = \frac{f^{(n+1)}(a)}{g^{(n+1)}(a)} .$$

16. In the light of Number 15 calculate the following limits.

(a) $\lim\limits_{x\sim 1} \frac{x^n - 1}{x - 1}$

(b) $\lim\limits_{x\sim 1} \frac{1 - x}{\log x}$

(c) $\lim\limits_{x\sim 0} \frac{\sin ax}{x}$

(d) $\lim\limits_{t\sim 0} \frac{\tan x - x}{x - \sin x}$

(e) $\lim\limits_{x\sim 0} \frac{x - \sin x}{\sin^3 x}$

(f) $\lim\limits_{x\sim \pi} (x - \pi)\tan \frac{x}{2}$

(g) $\lim\limits_{x\sim a} \frac{\sin x - \sin a}{x - a}$

(h) $\lim\limits_{x\sim 1} \{1 - \frac{1}{\log x} + \frac{1}{x - 1}\}$

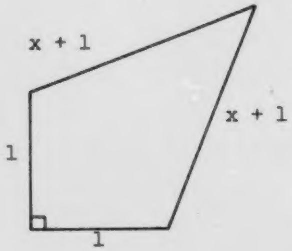
(i) $\lim\limits_{x\sim 0} \frac{1}{x} \log \frac{1 - \alpha x}{1 - \beta x}$

17. If $f - g$ has a zero of order n at a we say that $g(x)$ approximates $f(x)$ in the neighborhood of $x = a$ with an error of order n . We also say $f(x) = g(x) + A(x - a)^n$ plus terms of higher order (here $A = \frac{g^{(n)}(a) - f^{(n)}(a)}{n!}$) .

(a) Let s_1 , s_2 , s_3 , s_4 be successive sides of a convex quadrilateral. An ancient Egyptian document gives as the formula for the area of the quadrilateral

$$\frac{1}{4}(s_1 + s_3)(s_2 + s_4) .$$

This formula is correct for rectangles but is not generally valid. For a quadrilateral with two adjacent perpendicular sides of length 1 and two other sides of length $1 + x$, (see figure). What is the order of the error of the Egyptian formula in the neighborhood of $x = 0$?



(b) Let s be the arclength measured from X_0 to X along a plane curve. Determine the order in s to which the arclength is approximated by the chord length $\ell = |X - X_0|$ and give the error to lowest order.

13-4. Numerical Integration.

(i) The Rectangle Rule.

In Chapter 6 it was stated that the integral of a function can be defined as the limit of Riemann sums. This idea can be used directly to estimate the integral, but we shall see that it is easy to refine the idea so that with the same data and with little extra computation we may obtain much better estimates of the integral. For a function f defined on $[a,b]$ the given data comprise the values of the function at the points of a partition $\{x_0, x_1, \ldots, x_n\}$ of $[a,b]$. For simplicity, we require that the partition points be uniformly spaced: $x_k = a + kh$, $(k = 0,1,2,\ldots,n)$, where $h = \frac{b-a}{n}$. Let the function values at the partition points be $y_k = f(x_k)$.

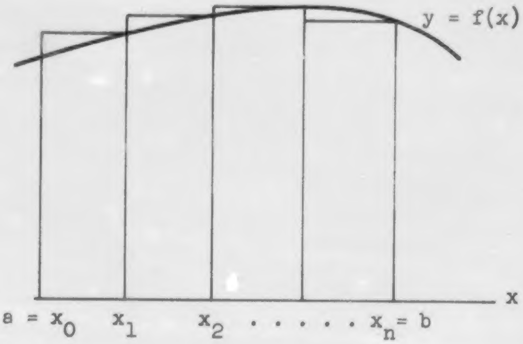


Figure 13 - 4a

We may approximate $\int_a^b f(x)dx$ by the Riemann sum

$$\sum_{k=1}^{n} y_k(x_k - x_{k-1}) = h\sum_{k=1}^{n} y_k .$$

This mode of approximation (the so-called Rectangle Rule) is tantamount to approximation of the integrand by the piecewise constant function $f^* : x \to f(x_k)$, for $x_{k-1} < x \leq x_k$. In order to estimate the error of the approximation, we use Taylor's Theorem. For the approximation f^* we have over one interval of the subdivision

$$\int_{x_{k-1}}^{x_k} f^*(x)dx = h\, y_k .$$

For the exact integrand, we have

$$\int_x^{x_k} f(\xi)d\xi = f(x_k)(x_k - x) + Q_1(x)$$

where $|Q_1(x)| < \frac{M_1}{2}(x_k - x)^2$; here M_1 is taken as an upper bound for $|f''(x)|$ on $[a,b]$. Replacing x by x_{k-1} in this result we have

$$\int_{x_{k-1}}^{x_k} f(x)dx = h\, y_k + \epsilon_k$$

where ϵ_k is quadratic or higher order in h,

$$|\epsilon_k| \le \frac{M_1 h^2}{2} .$$

Summing from 1 to n we obtain, finally

$$\text{(1a)} \qquad \int_a^b f(x)dx = h \sum_{k=1}^{n} y_k + \epsilon ,$$

where $\epsilon = \sum_{k=1}^{n} \epsilon_k$ satisfies $|\epsilon| \le \sum_{k=1}^{n} |\epsilon_k| \le \frac{nM_1 h^2}{2}$ or, since $h = \frac{b - a}{n}$,

$$\text{(1b)} \qquad |\epsilon| \le \frac{(b - a)^2 M_1}{2n} .$$

(ii) The Trapezoid Rule.

In contemplating the foregoing procedure we observe first that it does not make use of all the given data (the value y_0 does not appear in (1a)) and also to approximate the graph of f on a subinterval by a horizontal line seems rather crude. Instead we now approximate the graph by its chord (Figure 13-4(b)) on each subinterval of the partition; that is, we approximate f by the piecewise linear function f^* given by

$$f^*(x) = y_{k-1} + \frac{y_k - y_{k-1}}{h}(x - x_{k-1}) ,$$

for $x \in [x_{k-1}, x_k]$. For the integral of f^* we have the area of the "trapezoidal" region under the chord:

$$\text{(2)} \qquad \int_{x_{k-1}}^{x_k} f^*(x)dx = \frac{h}{2}(y_{k-1} + y_k)$$

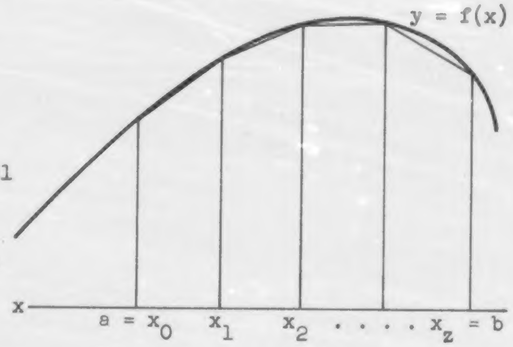


Figure 13-4b

To estimate the error of approximation in (2) we form the Taylor expansion of the original integral at x_{k-1}

$$(3)\quad \int_{x_{k-1}}^{x_k} f(t)dt = y_{k-1}(x_k - x_{k+1}) + \frac{y'_{k-1}}{2}(x_k - x_{k-1})^2 + \frac{y''_{k-1}}{6}(x_k - x_{k-1})^3 + \dots$$

$$= y_{k-1}h + \frac{y'_{k-1}}{2} h^2 + \frac{y''_{k-1}}{6} h^3 + \dots ,$$

where $y_m^{(n)} = f^{(n)}(x_m)$. Next we insert the Taylor expansion

$$y_k = y_{k-1} + y'_{k-1}\, h + \frac{y''_{k-1}}{2} h^2 + \dots$$

in (2) to obtain

$$(4)\qquad \int_{x_{k-1}}^{x_k} f^*(x)dx = y_{k-1}h + \frac{y'_{k-1}}{2} h^2 + \frac{y''_{k-1}}{4} h^3 + \dots .$$

From (3) and (4), on neglecting terms of order higher than 3 , we get the error estimate

$$(5)\qquad \epsilon_k = \int_{x_{k-1}}^{x_k} f^*(x)dx - \int_{x_{k-1}}^{x_k} f(x)dx$$

$$\approx \frac{y''_{k-1}\, h^3}{12} .$$

With the aid of the integral form for the Taylor remainder (Exercises 13-3, No. 9(a)) it can be proved that

$$(6)\qquad |\epsilon_k| \le \frac{M_2 h^3}{12}$$

where M_2 is an upper bound for $|f''(x)|$ on $[a,b]$; the proof is left to Exercises 13-4, Number 5(a). Summing, we obtain from (2), (5), and (6)

$$(7a)\qquad \int_a^b f(x)dx = h\Big(\frac{y_0}{2} + y_1 + y_2 + \dots + y_{n-1} + \frac{y_n}{2}\Big) + \epsilon ,$$

where $\epsilon = \sum_{k=1}^{n} \epsilon_k$; this is the Trapezoid Rule for numerical integration. From (6) we have

$$|\epsilon| \le \sum_{k=1}^{n} |\epsilon_k| \le \frac{nM_2h^3}{12} ;$$

whence

$$|\epsilon| \leq \frac{M_2(b - a)^3}{12n^2} . \tag{7b}$$

For sufficiently large n we see that the trapezoid rule will represent a considerable improvement over the Rectangle Rule given in (1a).

(iii) Simpson's Rule.

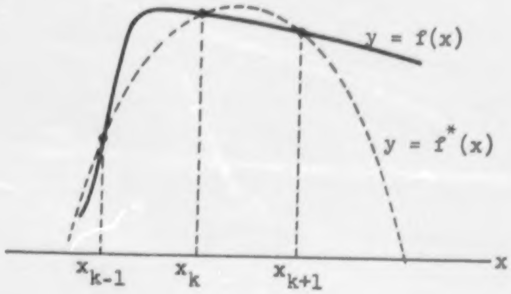


Figure 13-4c

Since approximation to the integrand by a piecewise linear function will usually be superior to approximation by a piecewise constant function it is natural to proceed further and see what improvement will be gained if the integrand is approximated by a piecewise quadratic function. To determine the three coefficients of a quadratic polynomial we need three conditions. Thus we use quadratic polynomials which take on the same value as f at three successive partition points to approximate f on two successive intervals of the subdivision (Figure 13-4c). For that reason we subdivide $[a,b]$ into an even number of subintervals. On the successive pairs of subintervals we approximate f by

$$f^* : x \to a_k(x - x_k)^2 + b_k(x - x_k) + c_k ,$$

for $x_{k-1} \leq x \leq x_{k+1}$, where k is odd. From the conditions

$$f^*(x_{k-1}) = y_{k-1} , f^*(x_k) = y_k , f^*(x_{k+1}) = y_{k+1}$$

we have

$$\left\{ \begin{aligned} a_k &= (y_{k-1} - 2y_k + y_{k+1})/2h^2 , \\ b_k &= (y_{k+1} - y_{k-1})/2h , \\ c_k &= y_k . \end{aligned} \right. \tag{8}$$

Consequently,

$$\int_{x_{k-1}}^{x_{k+1}} f^*(x)dx = \int_{x_k-h}^{x_k+h} [a_k(x - x_k)^2 + b_k(x - x_k) + c_k]dx ,$$

whence, from (8),

$$\int_{x_{k-1}}^{x_{k+1}} f^*(x)dx = \frac{h}{3}[y_{k-1} + 4y_k + y_{k+1}] \tag{9}$$

(compare Exercises 6-M, No. 11).

To estimate the error of approximation we use Taylor expansions at x_k. We write the original integral in the form

$$\int_{x_{k-1}}^{x_{k+1}} f(x)dx = \int_{x_k-h}^{x_k+h} [y_k + y_k'(x - x_k) + \frac{y_k''}{2}(x - x_k)^2 + \frac{y_k'''}{6}(x - x_k)^3 + \frac{y_k^{iv}}{24}(x - x_k)^4 + \dots]dx$$

to obtain

$$\int_{x_{k-1}}^{x_{k+1}} f(x)dx = 2y_kh + \frac{y_k''}{3}h^3 + \frac{y_k^{iv}}{60}h^5 + \dots . \tag{10}$$

For the function values y_{k-1} and y_{k+1} we have the Taylor expansions

$$y_{k-1} = y_k - y_k' h + \frac{y_k''}{2}h^2 - \frac{y_k'''}{6}h^3 + \frac{y_k^{iv}}{24}h^4 + \dots$$

and

$$y_{k+1} = y_k + y_k' h + \frac{y_k''}{2}h^2 - \frac{y_k'''}{6}h^3 + \frac{y_k^{iv}}{24}h^4 + \dots .$$

From these expansions, we obtain

$$\frac{h}{3}[y_{k-1} + 4y_k + y_{k+1}] = 2y_kh + \frac{y_k''}{3}h^3 + \frac{y_k^{iv}}{36}h^5 + \dots . \tag{11}$$

Comparing (10) and (11) we obtain the error estimate

$$\begin{aligned} \epsilon_k &= \int_{x_{k-1}}^{x_{k+1}} f^*(x)dx - \int_{x_{k-1}}^{x_{k+1}} f(x)dx \\ &\approx \frac{y_k^{iv}}{90}h^5 , \end{aligned} \tag{12}$$

where terms of order higher than 5 are neglected. In fact, it can be shown that

$$(13) \qquad |\epsilon_k| \leq \frac{M_4 h^5}{90}$$

where M_k is an upper bound for $|f^{IV}(x)|$ on $[a,b]$, (Exercises 13-4, No. 5(b)). Summing over the $\frac{n}{2}$ successive pairs of subintervals we then obtain Simpson's Rule,

$$(14a) \quad \int_a^b f(x)dx = \frac{h}{3}[y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \ldots + 2y_{n-2} + 4y_{n-1} + y_n] + \epsilon \,, \qquad (n \text{ even})$$

where, from (13) we have the estimate for the error

$$|\epsilon| \leq \sum_{k \text{ odd}} |\epsilon_k| \leq \frac{n}{2} \frac{M_4 h^5}{90} \,,$$

whence,

$$(14b) \qquad |\epsilon| \leq \frac{M_4 (b - a)^5}{180 \, n^4} \,.$$

Formula (14b) reveals a remarkable fact. If the integrand $f(x)$ were a quadratic polynomial we would expect zero error because we used quadratic interpolation to approximate f. Note, in addition, that if f is a polynomial function of degree 3 then f^{IV} is zero and we may take $M_4 = 0$ in (14b). Simpson's Rule yields more than we might expect, it is exact for third degree polynomials.

Evidently, we could go beyond Simpson's Rule and use interpolation polynomials of higher degree than two to approximate the integrand.

An interesting theory extends the methods developed in this section and it may be found in most of the introductory texts on numerical analysis.*

Example 13-4. Now let us compare the methods we have developed by approximating $\log 2 = \int_1^2 \frac{1}{x} dx$. For simplicity we subdivide the base interval $[1,2]$ into only two equal parts, by means of the partition $\{1, \frac{3}{2}, 2\}$. We have $y_0 = 1$, $y_1 = \frac{2}{3}$, $y_2 = \frac{1}{2}$. Applying the Rectangle Rule (1a) we obtain

---

*For example, see Henrici, P., Elements of Numerical Analysis, Wiley, New York, 1964.

$$\log 2 \approx \frac{1}{2}\left(\frac{2}{3} + \frac{1}{2}\right) \approx .583 \ ;$$

from the Trapezoid Rule (7a), we have

$$\log 2 \approx \frac{1}{2}\left(\frac{1}{2} + \frac{2}{3} + \frac{1}{4}\right) \approx .708 \ ;$$

finally, Simpson's Rule (14a) yields

$$\log 2 \approx \frac{1}{6}\left(1 + \frac{8}{3} + \frac{1}{2}\right) \approx .694 \ .$$

From a table of natural logarithms we have, to five decimal place accuracy $\log 2 = .69315$. From (14b) we estimate the error in Simpson's Rule using $D^4\left(\frac{1}{x}\right) = \frac{4!}{x^5} \leq 24$ for $x \in [1,2]$ and find

$$|\epsilon| \leq \frac{24}{180 \times 2^4} \leq \frac{1}{120} \ .$$

This error estimate is plainly too large. A more realistic estimate of the error is obtained if (12) is used for the error calculation instead of (13). Taking $y_0^{iv} = \frac{24}{(3/2)^5} \approx 3.2$ in (12), we obtain $\epsilon \approx \frac{1}{900}$. Although the use of (12) generally gives a better estimate of the error than (13), the method gives no clue as to whether the error is over - or underestimated.

∧ (iv) Stirling's Formula.

We have already taken note of the prodigious rate of growth of $n!$. The computation of $n!$ as a product, $1 \cdot 2 \cdot 3 \cdot 4 \cdots n$, also grows in difficulty extremely rapidly with n . Earlier, we established an upper bound for $n!$ ((13) of Section 8-6) which can be calculated easily with the help of a table of logarithms. Using the idea of numerical integration we can refine the estimate we have already obtained to get an asymptotic approximation to $n!$, the remarkable asymptotic formula of Stirling.

In order to approximate $n!$ using integrals we first convert the product to a sum by taking the logarithm:

$$\log(n!) = \log 1 + \log 2 + \ldots + \log n \ .$$

The form of the sum suggests approximation by the integral

$$(15) \qquad I_n = \int_1^n \log x \, dx = n \log n - n + 1 \ .$$

Since $\log x$ is increasing the integral is bounded by lower and upper Riemann sums as follows,

$$\sum_{k=1}^{n-1} \log n \le \int_1^n \log x \, dx \le \sum_{k=2}^{n} \log n \ .$$

From (15) we then have

$$\log(n - 1)! \le n \log n - n + 1 \le \log n! \ ,$$

whence

$$(16) \qquad e\left(\frac{n}{e}\right)^n \le n! \le en\left(\frac{n}{e}\right)^n \ .$$

This is a somewhat sharper result than (13) of Section 8-6.

We refine (16) by using the Trapezoid Rule. Since the graph $y = \log x$ is flexed downward, its chords lie below the graph, and the Trapezoid Rule gives an underestimate:

$$I_n^* = \frac{\log 1}{2} + \log 2 + \log 3 + \dots + \log(n - 1) + \frac{\log n}{2}$$

$$= \log n! - \frac{\log n}{2} \le \int_1^n \log x \, dx \ .$$

Consequently,

$$(17) \qquad n! \le e\sqrt{n}\left(\frac{n}{e}\right)^n \ .$$

In a similar fashion it is possible to obtain a lower estimate for $n!$ of the same form, namely, $c\sqrt{n}\left(\frac{n}{e}\right)^n$; this is left to Exercises 13-4, Number 6.

From these clues we suspect that there is a constant λ such that

$$\lim_{n\to\infty} \frac{n!}{\lambda\sqrt{n}\left(\frac{n}{e}\right)^n} = 1 \ .$$

We shall soon prove that such a constant exists, but for the moment let us take the existence of λ as a hypothesis and calculate its value. For the calculation we use Wallis's Product (see Example 10-6e) for π, namely

$$\pi = \lim_{n\to\infty} \frac{2}{2n + 1}\left[\frac{2^{2n}(n!)^2}{(2n)!}\right]^2$$

$$= \lim_{n\to\infty} \frac{2}{2n + 1}\left[\frac{2^{2n}\{\lambda\sqrt{n}\left(\frac{n}{e}\right)^n\}^2}{\sqrt{2n}\left(\frac{2n}{e}\right)^{2n}}\right]^2$$

$$= \lim_{n\to\infty} \frac{2}{2n + 1}\left[\frac{\lambda\sqrt{n}}{\sqrt{2}}\right]^2$$

$$= \frac{\lambda^2}{2} \ .$$

Consequently, if such a constant λ exists, then $\lambda = \sqrt{2\pi}$. Thus we obtain

(18a) $$\lim_{n \sim \infty} \frac{n!}{\sqrt{2\pi n}\,(\frac{n}{e})^n} = 1 .$$

Written in the form

(18b) $$n! \sim \sqrt{2\pi n}\,(\frac{n}{e})^n ,$$

this result is known as Stirling's Formula[*] and we say that $\sqrt{2\pi n}\,(\frac{n}{e})^n$ is an asymptotic expression for $n!$. For large n it is much easier to estimate $n!$ by the asymptotic formula than to calculate the product $1 \cdot 2 \cdot 3 \cdots (n - 1)n$ directly.

Next we shall prove the existence of a constant λ which satisfies (17). For this we set $\lambda_n = \dfrac{n!}{\sqrt{n}\,(\frac{n}{e})^n}$ and verify the existence of the limit $\lim_{n \sim \infty} \lambda_n = \lambda$. For the proof we shall show that λ_n is a bounded monotone function of n from which, by Lemma A10a, page 661, the result follows. From (17) we immediately obtain e as an upper bound for λ_n . Consequently $0 < \lambda_n < e$ and we have only to prove that λ_n is monotone. For this purpose we shall need an estimate of the error in the Trapezoid Rule. Set

$$J_k = \int_k^{k+1} \log x \, dx$$

and

$$J_k^* = \frac{\log k + \log(k + 1)}{2} .$$

We already know that $J_k > J_k^*$, and, since $D^2 \log x = -\frac{1}{x^2}$, we conclude from (6) that

(19) $$0 < J_k - J_k^* < \frac{1}{12\,k^2} .$$

Now, for

---

[*] James Stirling. English (1692-1770).

$$I_n = \sum_{k=1}^{n-1} J_k = \int_1^n \log x \, dx ,$$

$$I_n^* = \sum_{k=1}^{n-1} J_k^* = \log 2 + \log 3 + \dots + \frac{\log n}{2} ,$$

we have

$$I_n - I_n^* = \log \frac{e\sqrt{n}\,(\frac{n}{e})^n}{n!} = \log \frac{e}{\lambda_n} ,$$

and, from (19),

$$0 < \log \frac{e}{\lambda_n} < \frac{1}{12} \sum_{k=1}^{n-1} \frac{1}{k^2} .$$

Observe that

$$\begin{aligned} (20) \qquad \log \frac{\lambda_{n+1}}{\lambda_n} &= \log \frac{e}{\lambda_n} - \log \frac{e}{\lambda_{n+1}} \\ &= (I_n - I_n^*) - (I_{n+1} - I_{n+1}^*) \\ &= (I_{n+1}^* - I_n^*) - (I_{n+1} - I_n) \\ &= J_{n+1}^* - J_{n+1} . \end{aligned}$$

It follows from (19) that

$$(21) \qquad 0 < \log \frac{\lambda_n}{\lambda_{n+1}} < \frac{1}{12n^2} .$$

and from the left side of this inequality, that

$$\lambda_{n+1} < \lambda_n .$$

Since λ_n is bounded and monotone the existence of $\lim_{n\sim\infty} \lambda_n = \lambda$ is proved.

The estimate (21) can be used to refine Stirlings Formula. First, since λ_n is a decreasing function of n it follows that $\lambda = \lim_{n\sim\infty} \lambda_n < \lambda_n$,

From $\log \frac{\lambda_\nu}{\lambda_{\nu+1}} < \frac{1}{12\nu^2}$ we obtain, on summing from n to $n + k - 1$

(22) $$\log \frac{\lambda_n}{\lambda_{n+k}} < \frac{1}{12}\left[\frac{1}{n^2} + \frac{1}{(n+1)^2} + \dots + \frac{1}{(n+k-1)^2}\right].$$

But now observe that the expression in brackets is a lower Riemann sum for $\int_{n-1}^{n+k-1} \frac{1}{x^2}\,dx$, and consequently

$$\sum_{\nu=n}^{n+k-1} \frac{1}{x^2} < \int_{n-1}^{n+k-1} \frac{1}{x^2}\,dx < \int_{n-1}^{\infty} \frac{1}{x^2}\,dx \leq \frac{1}{n} .$$

From (22), then, for all k ,

$$\log \frac{\lambda_n}{\lambda_{n+k}} < \frac{1}{12(n-1)} .$$

Since $\lim_{k\sim\infty} \lambda_{n+k-1} = \lambda$, we conclude that

$$\log \frac{\lambda_n}{\lambda} \leq \frac{1}{12(n-1)}$$

or

$$\lambda_n \leq \lambda \exp \{\frac{1}{12(n-1)}\} .$$

Consequently, $n! = \lambda_n \sqrt{n}\, (\frac{n}{e})^n < \sqrt{2\pi n}\, (\frac{n}{e})^n e^{1/12(n-1)}$. In summary, we have found

(23) $$\sqrt{2\pi n}\, (\frac{n}{e})^n < n! < \sqrt{2\pi n}\, (\frac{n}{e})^n e^{1/12(n-1)} .$$

Exercises 13-4

1. (a) Estimate π by approximation to

$$\frac{\pi}{4} = \int_0^1 \frac{1}{1+x^2}\,dx .$$

(b) Estimate π by approximation to

$$\frac{\pi}{6} = \int_0^{1/2} \frac{1}{\sqrt{1-x^2}}\,dx .$$

(c) Estimate how large n should be taken in Simpson's Rule to give π accurately to 5 places by approximation to the integral of Part (a).

2. Obtain $\log 3$ to four decimal place accuracy by numerical integration of $\int_1^3 \frac{1}{x}\,dx$.

3. Estimate the integral

$$\int_0^{\pi/2} \frac{d\psi}{\sqrt{\cos \psi}}$$

of Section 13-1, Equation (1). (Hint: Compare Exercises 13-1, Number 2. Use the substitutions $\sin \psi = u^2$, $\cos \psi = 1 - \frac{v^2}{2}$ to obtain regular algebraic integrals.)

4. Verify that

$$\int_0^4 (x - 4)(x - 2)x\,dx = 0 .$$

5. Use the integral form of the Taylor remainder to obtain the error bounds for

(a) the Trapezoid Rule given in Formula (6) and

(b) Simpson's Rule given by Formula (13).

6. Using approximation to the integral $\int_1^n \log x\,dx$ obtain an inequality of the form

$$c\sqrt{n}\,(\tfrac{n}{e})^n \leq n!$$

(Hint: Note that the extension to the left of a chord to the graph $y = \log x$ lies above the curve.)

7. Obtain asymptotic expressions for the following binomial coefficients.

(a) $\binom{2n}{n}$

(b) $\binom{n}{k}$, k fixed.

(c) $\binom{pn}{n}$, for large p and n .

8. Obtain an asymptotic expression for the coefficient of x^{2n+1} in the Taylor expansion of arc sin x (given in Example 13-3b).

9. Obtain a sharper lower bound for $J_k - J_k^*$ than that of (19) and so improve the lower estimate for n! in (23). (Hint: Use the integral form for the Taylor remainder as in No. 5.)

∧13-5. Numerical Solution of First Order Differential Equations.

Picard's method (Section 13-2) for the solution of the initial value problem for the first order differential equation

$$\frac{dy}{dx} = \Phi(x,y) \tag{1}$$

is used mainly to prove existence and uniqueness of the solution. Practical numerical methods usually proceed in an entirely different fashion. Here we give one of the simplest techniques, Euler's method. There are many more sophisticated methods for the solution of (1) and new methods for the solution of differential equations are continually being devised and investigated, but Euler's method will suffice as an introduction.

In order to obtain error estimates by the means at our disposal we shall consider the special case of separable equations

$$\frac{dy}{dx} = f(x)g(y) \ . \tag{2a}$$

With minor adaptations our arguments can be extended to the general case. Our problem will be to obtain a numerical solution of (2a) on the interval $[a,b]$ subject to the initial condition

$$y = y_0 \quad \text{at} \quad x = a \ . \tag{2b}$$

By a numerical solution we mean a tabulation $y_0 , y_1 , y_2 , \dots , y_n$ which gives the values of y at the respective points $x_0 , x_1 , x_2 , \dots , x_n$, of a partition of $[a,b]$. We shall suppose that the partition points are uniformly spaced; $x_k = x_0 + kh$, for $k = 0 , 1 , 2 , \dots , n$, where $h = \frac{b - a}{n}$. Of course, we do not expect to obtain y_k exactly but to get an approximation $\hat{y}_k$. We shall then estimate the error $\epsilon_k = |\hat{y}_k - y_k|$.

First, we observe for the exact solution that

$$y_k = y_{k-1} + y'_{k-1} h + y''_{k-1} \frac{h^2}{2} + \dots , \tag{3}$$

where $y'_k = \Phi(x_k,y_k)$. This suggests that we obtain the approximate solution stepwise by

$$(4)\qquad \begin{cases} \hat{y}_1 = y_0 + h\,\Phi(x_0,y_0) \\ \hat{y}_2 = \hat{y}_1 + h\,\Phi(x_1,\hat{y}_1) \\ \quad\cdots \\ \hat{y}_n = \hat{y}_{n-1} + h\,\Phi(x_{n-1},\hat{y}_{n-1}). \end{cases}$$

We may interpret $\Phi(x,y)$ as a function which assigns a slope at each point of some region of the plane. A solution of the differential equation (1) is then a curve $y = F(x)$ which has at each of its points $(x,F(x))$ the slope $\Phi(x,F(x))$. The system of Equations (4) then defines the successive vertices of an approximating polygon (the Euler polygon) which begins at the initial vertex (x_0,y_0) and continues along the line of slope $\Phi(x_0,y_0)$ until $x = x_1$, where the next vertex $(x,\hat{y}_1)$ is reached, and so on (Figure 13-5).

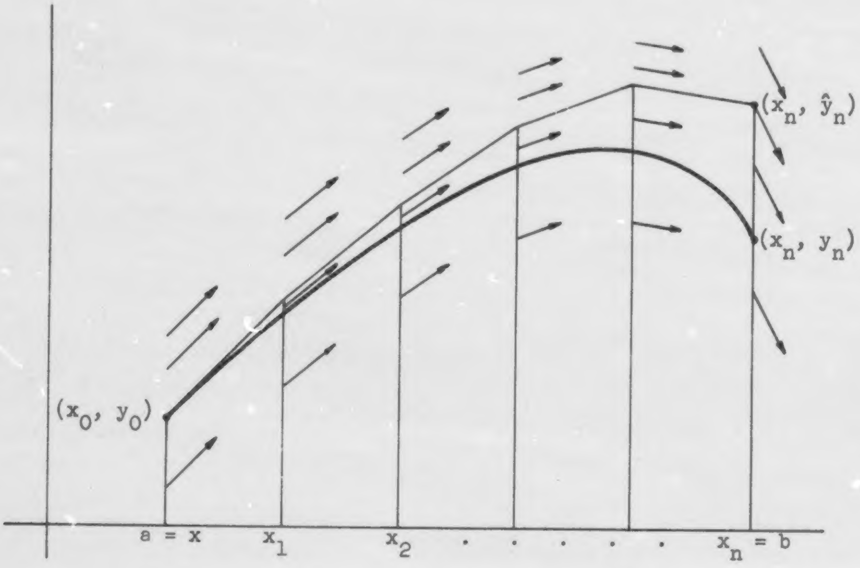


Figure 13-5

We shall prove that the approximation can be kept within any prescribed tolerance provided the norm h of the partition is made small enough*

---

[*] It is important not to assume without prior analysis that making the norm small will always reduce the error for all "difference schemes" similar to the Euler method. Counterexamples, by no means artificially contrived, are known. (See Isaacson, E. and Keller, H., Analysis of Numerical Methods, Wiley. New York 1966).

From (3) and (4) we have the error

$$(5) \qquad \hat{y}_k - y_k = (\hat{y}_{k-1} - y_{k-1}) + h[\,\Phi(x_{k-1},\hat{y}_{k-1}) - y'_{k-1}] - \frac{h^2}{2} y''_{k-1} + \cdots .$$

For the coefficient of the first order term we have by the Law of the Mean

$$\begin{aligned}\Phi(x_{k-1},\hat{y}_{k-1}) - y'_{k-1} &= f(x_{k-1})[g(\hat{y}_{k-1}) - g(y_{k-1})] \\ &= f(x_{k-1})g'(\eta_k)(\hat{y}_{k-1} - y_{k-1}) ,\end{aligned}$$

where η_k is some number between y_{k-1} and $\hat{y}_{k-1}$. It follows from (5), that

$$(6) \qquad \hat{y}_k - y_k = (\hat{y}_{k-1} - y_{k-1})[1 + hf(x_{k-1})g'(\eta_k)] - \frac{h^2}{2} y''_{k-1} + \cdots .$$

On the assumption that f and g have bounded derivatives the coefficients of h in (6) are bounded and it is possible to find nonnegative constant A (see Exercises 13-5, No. 1) such that for the absolute error $\epsilon_k = |\hat{y}_k - y_k|$ we have

$$(7) \qquad \epsilon_k \le \epsilon_{k-1}(1 + hA) + \frac{h^2}{2} A .$$

We now use (7) to express ϵ_{k-1} in terms of ϵ_{k-2} , ϵ_{k-2} in terms of ϵ_{k-3} , etc., and obtain

$$\begin{aligned}\epsilon_k &\le \epsilon_{k-2}(1 + hA)^2 + \frac{h^2}{2} A[1 + (1 + hA)] \\ &\le \epsilon_{k-3}(1 + hA)^3 + \frac{h^2}{2} A[1 + (1 + hA) + (1 + hA)^2] \\ &\quad \cdots \\ &\le \epsilon_0(1 + hA)^k + \frac{h^2}{2} A[1 + (1 + hA) + \ldots + (1 + hA)^{k-1}] .\end{aligned}$$

Summing the geometric progression in brackets we then have

$$\epsilon_k \le \frac{h}{2}(1 + hA)^k ,$$

where we have taken $\hat{y}_0 = y_0$ so that $\epsilon_0 = 0$. We put this last inequality in the form

$$\epsilon_k \le \frac{b - a}{2n}\left[1 + \frac{(b - a)A}{n}\right]^k .$$

Since, by Section 8-6, Equation (11),

$$[1 + \frac{(b-a)A}{n}]^{\frac{n}{(b-a)A}} \leq e \; ,$$

we obtain

$$\epsilon_k \leq \frac{b-a}{2n} \exp\{(b-a)A \frac{k}{n}\} \tag{8}$$

$$\leq \frac{b-a}{2n} \exp\{(b-a)A\} \; .$$

From (8), we conclude that $\lim_{n\sim\infty} \epsilon_k = 0$.

The only source of error we have made allowance for is the truncation error caused by the cut-off of the Taylor expansion after the first order term. There will also be round-off error from the approximate computation of $\hat{y}_k$. We cannot treat round-off error casually because any error committed at one step of the computation is propagated through all succeeding steps. At the same time, it is wasteful to calculate $\hat{y}_k$ to a level of accuracy much higher than the inherent truncation error since such accuracy will not significantly affect the final result. Thus the round-off error should be comparable to the truncation error. In (7) the truncation error is bounded by $\frac{h^2}{2} A$. To allow for round-off error as well we need only bound the round-off error similarly and replace A by a larger constant. In that case we obtain the same kind of final estimate as (8). If we do not take the precaution of reducing the round-off error appropriately, we may lose any benefit from refining the partition of $[a,b]$.

Example 13-5. We consider the equation of motion for a pendulum, Section 12-3, Equation (17), with amplitude $\frac{\pi}{2}$. For simplicity we take $\frac{g}{2\ell} = 1$ and obtain the equation for the first quarter period,

$$\frac{d\theta}{dt} = \sqrt{\cos\theta} \qquad 0 \leq \theta \leq \frac{\pi}{2} \tag{9a}$$

with the initial condition

$$\theta = 0 \text{ at } t = 0 \; . \tag{9b}$$

To put a bound on the error we first replace (3) by the Taylor Formula with remainder

$$\theta_{k+1} = \theta_k + \theta_k' h + R_{2,k} \tag{10}$$

where $R_{2,k} \leq M_2 \frac{h^2}{2}$. Here M_2 is an upper bound for θ_k'' on the domain of interest. From (9a) we have within the first quarter cycle of the motion $(0 \leq \theta \leq \frac{\pi}{2})$,

$$\theta'' = \frac{d^2\theta}{dt^2} = (\frac{d}{d\theta}\sqrt{\cos\theta})\frac{d\theta}{dt}$$

$$= -\frac{1}{2}\sin\theta \, ;$$

thus we take $M_2 = \frac{1}{2}$. Now the error in the Euler method is obtained like (6) in the form

$$\hat{\theta}_k - \theta_k = (\hat{\theta}_{k-1} - \theta_{k-1})\left[1 - \frac{h}{2}\frac{\sin\psi_{k-1}}{\sqrt{\cos\psi_{k-1}}}\right] - R_{2,k} + r_k \tag{11}$$

where ψ_{k-1} lies between $\hat{y}_{k-1}$ and y_{k-1} and r_k is the round-off error. We cannot use this formula for error estimates over an entire quarter cycle of the motion. It can be shown (Exercises 11-5, No. 2) that $\hat{\theta}_k \geq \theta_k$ for $k > 0$. The value $\frac{\pi}{2}$ will then be reached by the approximate solution earlier than by the exact solution. Thus in a neighborhood of $\theta = \frac{\pi}{2}$ $\sqrt{\cos\psi_{k-1}}$ may be arbitrarily close to zero and no bound can be put on $\frac{\sin\psi_{k-1}}{\sqrt{\cos\psi_{k-1}}}$. Nonetheless we may fix our attention on any value $0 < \theta^* < \frac{\pi}{2}$ and consider the error in $\hat{\theta}_k$ when $\hat{\theta}_{k-1} \leq \theta^*$. We have

$$\frac{\sin\psi_{k-1}}{\sqrt{\cos\psi_{k-1}}} < \frac{\sin\hat{\theta}_{k-1}}{\sqrt{\cos\hat{\theta}_{k-1}}} \leq \frac{\sin\theta^*}{\sqrt{\cos\theta^*}} \, .$$

Entering this estimate in (11) and using the estimate obtained for $R_{2,k}$ we obtain

$$\epsilon_k \leq \epsilon_{k-1}\left[1 - \frac{h}{2}\frac{\sin\theta^*}{\sqrt{\cos\theta^*}}\right] + \frac{h^2}{4} + |r_k| \, .$$

We bound the round-off error by $|r_k| \leq \frac{h^2}{4}$ and take θ^* close enough to $\frac{\pi}{2}$ so that $\frac{\sin\theta^*}{\sqrt{\cos\theta^*}} \geq 1$ and obtain by (8) with $A = \frac{\sin\theta^*}{\sqrt{\cos\theta^*}}$ and $(b - a) = \frac{\tau}{4}$ where τ is the period of the pendulum,

$$\epsilon_k \leq \frac{h}{2} e^{\tau A/4} \, . \tag{12}$$

The method does not permit us to determine τ, nor do we know beforehand how large k must be for $\hat{\theta}_k$ to exceed θ^*. However, we may make an independent determination of τ (Exercises 13-1, No. 2 and 13-4, No. 3) and knowing that $\theta = \frac{\pi}{2}$ at $t = \frac{\tau}{4}$ we can obtain an accurate description of the motion by Euler's method.

Exercises 13-5

1. Consider the solution by Euler's method of the initial value problem (2a,b) in a region where $f(x)$ and $g(y)$ have bounded derivatives. Obtain error estimates in the form of Equation (8).
2. Show for the initial value problem (9a,b) that the approximate solution is greater than the exact solution, namely that $\hat{\theta}_k > \theta_k$, $(k > 0)$.
3. Show if $\Phi(x,y)$ is a function of x alone, that Euler's method for Equation (1) approximates the solution by successive Riemann sums.

Miscellaneous Exercises

1. (a) Obtain an iteration scheme for the zero of $f : x \rightarrow a - \frac{1}{x}$ and thus show how to calculate the reciprocal of a without divisions.

   (b) Use the method obtained in (a) to calculate $\frac{1}{\pi}$ accurately to the extent indicated by the approximation $\pi \approx 3.141593$.

2. In Section 13-2 we observed if $x_0 > 0$ is an approximation to $\sqrt{A}$ from one side, then $\frac{A}{x_0}$ is an estimate from the other side. We showed (for $A = 7$, but the proof is valid in general) that the arithmetic mean $x_1 = \frac{1}{2}(x_0 + \frac{A}{x_0})$ is an approximation from above. Show that the harmonic mean approximates $\sqrt{A}$ from below and estimate the error.

3. Compute $\int_0^{\pi} \frac{\sin x}{x} dx$ accurately to three decimal places.

4. (a) Consider a right triangle with shorter side of length a , longer side b , and hypotenuse c . Let α be the angle opposite side a . Estimate the error in the approximation

$$\alpha \approx \frac{3a}{b + 2c} .$$

   (b) Obtain an approximation for α in the form

$$\alpha \approx \frac{a(pb + qc)}{c^2} ,$$

   where the constants p and q are so chosen that the error in the approximation is higher order than 3 . Estimate the error.

5. Consider the solid of revolution obtained by rotating the graph $y = f(x)$ of a nonnegative function $[a,b]$ about the x-axis. Let A_0 , A_1 , A_2 be the areas of the cross-sections of the solid perpendicular to the x-axis at $x = a$, $\frac{a + b}{2}$, b , respectively. Show that Simpson's Rule

$$V = \frac{b - a}{6}[A_0 + 4A_1 + A_2]$$

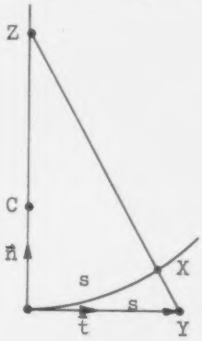
   gives the exact volume for each of the following cases,

   (a) frustrum of a right circular cone ($y = f(x)$ is a straight line),

   (b) segment of a sphere ($y = f(x)$ is an arc of a circle with center on the x-axis),

(c) segment of paraboloid, ellipsoid or hyperboloid of revolution ($y = f(x)$ is an arc respectively of parabola, ellipse or hyperbola, with the x-axis as an axis of symmetry).

Λ6. Let $\vec{X} = \vec{r}(s)$ be the vectorial representation of a plane curve with arclength as parameter. Let $\vec{0} = \vec{r}(0)$. Set $\vec{Y} = s\vec{t}$ where $\vec{t}$ is the tangent at 0. Let Z be the point where the line XY meets the normal line through $\vec{0}$. Show if the curvature κ at 0 is not zero then to lowest order in s, $\vec{Z} = 3\vec{C}$ where C is the center of curvature.



Λ7. Let a curve be given by $\vec{X} = \vec{r}(s)$ where s is arclength measured from $\vec{0} = \vec{r}(0)$. Consider any three distinct points $\vec{X}_1 = \vec{r}(s_1)$, $\vec{X}_2 = \vec{r}(s_2)$, $\vec{X}_3 = \vec{r}(s_3)$ where the s_i are confined to an δ-neighborhood of $s = 0$. Show that the circle through the three points approaches the osculating circle as δ approaches zero. (Assume the curvature at $s = 0$ is not zero.)

Chapter 14

SEQUENCES AND SERIES

14-1. Introduction.

Sequences of numbers are not new to us. In Section 1-2 we were concerned with the sequence $I_1, I_2, \ldots,$ of approximants to the area under a curve. At the beginning of Section 13-2 we investigated the sequence defined recursively by

$$a_1 = 3 \ , \ a_{k+1} = \frac{1}{2}\left(a_k + \frac{7}{a_k}\right)$$

as a sequence approximating $\sqrt{7}$. While the notion of a sequence as an infinite succession of numbers; first a_1, next a_2, and so on, marching off indefinitely is natural, we formally define a sequence as a function $a : k \to a_k$ whose domain $\{k\}$ is the set of natural numbers. The natural number k, the position of a_k in the sequence, is called the index, and a_k is called the k-th term of the sequence.

Our primary concern is whether a sequence serves to approximate a given real number ℓ, in the sense that to any given margin of error $\epsilon > 0$ there exists a number $\omega = \Omega(\epsilon)$ such that every term a_k with $k > \omega$ is within ϵ of ℓ, i.e. $|a_k - \ell| < \epsilon$ for $k > \omega$. Under these circumstances we write $\lim\limits_{k \sim \infty} a_k = \ell$. This merely repeats the general definition, p. 232, in Section 5-8 of $\lim\limits_{x \sim \infty} f(x) = \ell$ for sequences, in particular. Moreover, given a sequence $a : k \to a_k$ for which there is a number ℓ such that the statement "$\lim\limits_{k \sim \infty} a_k = \ell$" is true, we say that the sequence a "converges" and that it "converges to ℓ." If a sequence does not converge we say it diverges.

The two crucial problems are:

1. Given a sequence to determine whether or not it converges.
2. To determine the number to which the sequence converges.

Since for any given $\epsilon > 0$, we need only look at a_k with $k > \omega$ we need only insist that ω be greater than j if we want to ignore the terms $a_1, a_2, \ldots, a_j$. Thus, for questions of convergence the beginning of a sequence is irrelevant.

14-2. Convergence of Sequences.

We formally state the definition of $\lim\limits_{k \to \infty} a_k$ as

DEFINITION 14-2a. We write $\lim\limits_{k \to \infty} a_k = \ell$ if for every $\epsilon > 0$ there is a number $\omega = \Omega(\epsilon)$ such that for all $k > \omega$ the inequality

$$|a_k - \ell| < \epsilon$$

is satisfied.

This definition is identical with Definition 3-2 for $a = \infty$ if we define the deleted neighborhood of ∞ as a set $\{x : x > \omega\}$. Because of this identification all the theorems and proofs of Section 3-4 can be transtated in terms of sequences. Thus we simply state the theorems and leave their proofs to the exercises (Exercises 14-2, No. 1).

THEOREM 14-2a. For a constant sequence $a : k \to c$

$$\lim_{k \to \infty} a_k = c .$$

THEOREM 14-2b. If $\lim\limits_{k \to \infty} a_k = \ell$ then for any constant c,

$$\lim_{k \to \infty} ca_k = c \lim_{k \to \infty} a_k = c\ell .$$

THEOREM 14-2c. If $\lim\limits_{k \to \infty} a_k = \ell$ and $\lim\limits_{k \to \infty} b_k = m$, then

$$\lim_{k \to \infty} (a_k + b_k) = \lim_{k \to \infty} a_k + \lim_{k \to \infty} b_k = \ell + m$$

Corollary. The limit of a linear combination of sequences is the same linear combination of the limits of the sequences; i.e., if

$$\lim_{k \to \infty} a_k^{(i)} = \ell_i \quad , \quad i = 1 , 2 , \ldots , n$$

then

$$\lim_{k \to \infty} \sum_{i=1}^{n} c_i a_k^{(i)} = \sum_{i=1}^{n} c_i \lim_{k \to \infty} a_k^{(i)} = \sum_{i=1}^{n} c_i \ell_i .$$

THEOREM 14-2d. If $\lim\limits_{k\sim\infty} a_k = \ell$ and $\lim\limits_{k\sim\infty} b_k = m$,

then
$$\lim_{k\sim\infty} a_k\, b_k = \ell m \ .$$

Lemma 14-2. If $\lim\limits_{k\sim\infty} a_k = \ell$ and $\ell > 0$, then there exists a number ω such that $a_k > 0$ for $k > \omega$.

Corollary 1. If $\lim\limits_{k\sim\infty} a_k = \ell$ and $\ell \neq 0$, then there exists a number ω such that
$$\left|\frac{3\ell}{2}\right| > |a_k| > \left|\frac{\ell}{2}\right|$$
for $k > \omega$.

Corollary 2. A limit of a sequence whose values are nonnegative (nonpositive) is nonnegative (nonpositive).

THEOREM 14-2e. If $\lim\limits_{k\sim\infty} a_k = \ell \neq 0$, then
$$\lim_{k\sim\infty} \frac{1}{a_k} = \frac{1}{\ell} \ .$$

Corollary. If $\lim\limits_{k\sim\infty} a_k = \ell$ and $\lim\limits_{k\sim\infty} b_k = m \neq 0$

then
$$\lim_{k\sim\infty} \frac{a_k}{b_k} = \dots = \frac{\ell}{m} \ .$$

THEOREM 14-2f. If $a_k \leq b_k$ for $k > \omega$, and $\lim\limits_{k\sim\infty} a_k = \ell$ and $\lim\limits_{k\sim\infty} b_k = m$,

then $\ell < m$.

Corollary 1. (**Sandwich Theorem**) If
$$a_k \leq b_k \leq c_k \quad \text{for} \quad k > \omega \ ,$$
and if
$$\lim_{k\sim\infty} a_k = \ell \quad \text{and} \quad \lim_{k\sim\infty} c_k = n \ ,$$

then, if $\lim\limits_{k \sim \infty} b_k$ exists,

$$\ell \leq \lim_{k \sim \infty} b_k \leq n \ .$$

Corollary 2. (Squeeze Theorem) If $a_k \leq b_k \leq c_k$ for $k > \omega$ and if

$$\lim_{k \sim \infty} a_k = \lim_{k \sim \infty} c_k = \ell \ ,$$

then

$$\lim_{k \sim \infty} b_k = \ell \ .$$

Note that the hypothesis of the Squeeze Theorem forces $\lim\limits_{k \sim \infty} b_k$ to exist, while this must be part of the hypothesis of the Sandwich Theorem. For example, let $a : k \to -1$, $c : k \to 1$ and observe that $b : k \to (-1)^k$ does not converge.

Corollary. If c_n is between a_n and b_n for all n and $\lim\limits_{n \sim \infty} a_n = \lim\limits_{n \sim \infty} b_n$, then $\lim\limits_{n \sim \infty} c_n = \lim\limits_{n \sim \infty} a_n$.

THEOREM 14-2g. If $\lim\limits_{k \sim \infty} a_k = \ell$ and f is continuous at ℓ and the terms of the sequence is in the domain of f , then the sequence $b_k = f(a_k)$ converges to $f(\ell)$, i.e.,

$$\lim_{k \sim \infty} b_k = \lim_{k \sim \infty} f(a_k) = f(\lim_{k \sim \infty} a_k) = f(\ell) \ .$$

Proof.

(1) Since f is continuous at ℓ , corresponding to any $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - f(\ell)| < \epsilon$ if $|x - \ell| < \delta$ and x is in the domain of f.

(2) Since $\lim\limits_{k \sim \infty} a_k = \ell$, corresponding to the δ of 1 , there exists an $\omega = \Omega(\delta)$ such that $|a_k - \ell| < \delta$ for $k > \omega$.

(3) Since a_k is in the domain of f , combining 1 and 2 we have

$$|f(a_k) - f(\ell)| < \epsilon$$

for $k > \omega$.

With the observation that $\lim\limits_{k\sim\infty} \frac{1}{k} = 0$ (take $\Omega(\epsilon) = \frac{1}{\epsilon}$) and the aid of the preceding theorems we can find the limits of many important sequences.

Example 14-2a.

$$\lim_{n\sim\infty} \frac{a_k n^k + a_{k-1} n^{k-1} + \ldots + a_0}{b_j n^j + b_{j-1} n^{j-1} + \ldots + b_0} = \begin{cases} \frac{a_k}{b_j} & \text{if} \quad k = j \\ 0 & \text{if} \quad k < j \, , \end{cases}$$

since $$\frac{a_k n^k + \ldots + a_0}{b_j n^j + \ldots + b_0} = \frac{1}{n^{j-k}} \, \frac{a_k + a_{k-1}(\frac{1}{n}) + \ldots + a_0(\frac{1}{n})^k}{b_j + b_{j-1}(\frac{1}{n}) + \ldots + b_0(\frac{1}{n})^k} \, .$$

Example 14-2b. If $|r| < 1$ then $\lim\limits_{n\sim\infty} r^n = 0$. Let $\frac{1}{|r|} = 1 + p$, $p > 0$. Then $|r^n| = \frac{1}{(1 + p)^n} < \frac{1}{1 + np}$ (from Exercises A3-1, No. 6, or the Binomial Theorem, Exercises A3-2, No. 13(c)) but $\lim\limits_{n\sim\infty} \frac{1}{1 + np} = 0$ and the conclusion follows from the Squeeze Theorem.

Example 14-2c. If $a > 0$, then $\lim\limits_{n\sim\infty} \sqrt[n]{a} = 1$. Let $a > 1$ and set $\epsilon : n \to \epsilon_n = \sqrt[n]{a} - 1$. Hence $1 + \epsilon_n = \sqrt[n]{a}$ and $a = (1 + \epsilon_n)^n > 1 + n\epsilon_n$. Therefore $\epsilon_n < \frac{a - 1}{n}$. But $\lim\limits_{n\sim\infty} \frac{a - 1}{n} = 0$ and thus $\lim\limits_{n\sim\infty} \epsilon_n = 0$ by the Squeeze Theorem.

If $0 < a < 1$ then $$\lim_{n\sim\infty} \sqrt[n]{a} = \lim_{n\sim\infty} \frac{1}{\sqrt[n]{1/a}} = \frac{1}{\lim\limits_{n\sim\infty} \sqrt[n]{1/a}} = 1 \, .$$

Example 14-2d. $$\lim_{n\sim\infty} \sqrt[n]{n} = 1 \, .$$

Set $\sqrt[n]{n} = 1 + \epsilon_n$,

$$n = (1 + \epsilon_n)^n \geq 1 + n\epsilon_n + \frac{n(n - 1)}{2} \epsilon_n^2 \geq \frac{n(n - 1)}{2} \epsilon_n^2 > \frac{(n - 1)^2}{2} \epsilon_n^2$$

whence $\epsilon_n < \sqrt{\frac{2}{n - 1}}$, $(n > 1)$. Consequently,

$$\lim_{n\sim\infty} \sqrt[n]{n} = \lim_{n\sim\infty} (1 + \epsilon_n) = 1 \, .$$

Theorems 14a-g exhibit the following pattern: under certain conditions the limit of the sequence $a : k \to a_k$ is ℓ. In each case there is the implicit assertion that the sequence has a limit. We close this section with three existence theorems. Under appropriate conditions we shall guarantee the existence of the limit of a sequence without necessarily being able to determine its value.

THEOREM 14-2h. (Monotone Convergence Theorem) If the sequence $a : k \to a_k$ is bounded and nondecreasing, $a_k \leq a_{k+1}$ for all k, then the sequence a converges, and the limit is the least upper bound of the range of a.

Proof. Let the least upper bound of the range of a be written* as $\ell = \sup\{a_k\}$. Given any $\epsilon > 0$, there exists an integer $\nu = k_\epsilon$ satisfying $\ell - \epsilon < a_\nu \leq \ell$. Hence if $k > \nu = N(\epsilon)$ then $\ell - \epsilon < a_\nu \leq a_k \leq \ell$, or $|\ell - a_k| < \epsilon$.

Corollary. If the sequence $a : k \to a_k$, is nonincreasing and the range of a is bounded, then $\lim_{k\sim\infty} a_k = \inf\{a_k\}$.

Proof. Consider $b_k = -a_k$ and apply the preceding theorem.

Example 14-2e. The sequence $a : k \to a_k = (1 + \frac{1}{k})^k$ converges. To show that a is a monotone sequence, consider $f : t \to (1 + \frac{1}{t})^t$. Then

$$f'(t) = (1 + \frac{1}{t})^t(\log(1 + \frac{1}{t}) - \frac{1}{t+1}) .$$

But, by the Mean Value Theorem,

$$\log(1 + \frac{1}{t}) = \log(1 + \frac{1}{t}) - \log 1 = \frac{1}{t}\,\frac{1}{u} ,$$

where

$$1 < u < 1 + \frac{1}{t} .$$

Hence $\frac{1}{1 + \frac{1}{t}} < \frac{1}{u}$ and therefore $\log(1 + \frac{1}{t}) > \frac{1}{t}\,\frac{1}{1 + \frac{1}{t}} = \frac{1}{t+1}$.

---

*Recall from the footnote on p. 265 that the least upper bound of $\{a_k\}$ is called the supremum of $\{a_k\}$ and written $\sup\{a_k\}$. We shall also introduce the greatest lower bound or infinum, indicated by $\inf\{a_k\}$.

Hence f' is positive and f is increasing. Thus the sequence a is increasing. On the other hand, by the Binomial Theorem

$$(1 + \tfrac{1}{k})^k = \sum_{j=0}^{k} (\tfrac{1}{k})^j \binom{k}{j}$$

$$= \sum_{\ell=0}^{k} \frac{k.(k-1) \ldots (k-L+1)}{k.k \ldots k} \cdot \frac{1}{\ell!}$$

$$< \sum_{j=0}^{k} \frac{1}{j!} < 1 + 1 + \frac{1}{2} + \frac{1}{6} + \sum_{j=4}^{k} \frac{1}{2^j} ,$$

since $2^j < j!$ for $j > 4$. But, since $\sum_{j=4}^{k} \frac{1}{2^j} < \frac{1}{16} \sum_{j=0}^{k} \frac{1}{2^j} = \frac{1}{16} \frac{1 - (\frac{1}{2})^j + 1}{1 - \frac{1}{2}} < \frac{1}{8}$, we have $(1 + \frac{1}{k})^k < 3$. Hence, since a is nondecreasing and the range of a is bounded it follows, from Theorem 14-2h, that a converges. (This result can also be obtained by the techniques of p. 480 f.)

The idea of a subsequence obtained by the deletion of terms from a given sequence is frequently useful. A subsequence of $a : k \to a_k$ is defined by an increasing sequence of positive integers $i : k \to i_k$; the subsequence corresponding to i being $b : k \to b_k = a_{i_k}$.

THEOREM 14-2i. Every sequence possesses a monotone subsequence.

> This property of sequences is very general: the proof uses only one property of the set of real numbers R , that R is ordered; thus the theorem applies to any sequence whose range is contained in an ordered set.

Proof. We prove the sequence $a : k \to a_k$ contains either a nondecreasing subsequence or a decreasing subsequence. For the proof we assume that a does not have a nondecreasing subsequence and show that a must then have a decreasing subsequence.

Since a does not contain a nondecreasing sequence, it follows for any term a_k that any finite nondecreasing subsequence beginning with a_k must terminate, therefore there exists a subsequence of maximal length n .

(1) $$a_k = a_{j1} \leq a_{j2} \leq \cdots \leq a_{jn} , \quad (j_1 \leq j_2 \leq \cdots \leq j_n) .$$

It follows that $a_{jn+m} < a_{jn}$, for all terms a_{jn+m} following a_{jn}, otherwise the sequence could be extended to one more term by including a_{jn+m}. Now, to each k associate the index of a maximal sequence beginning with a_k; i.e., set $j_n = \nu(k)$ in (1). Then set $b_1 = a_{p_1}$, $b_2 = a_{p_2}$, ..., $b_k = a_{p_k}$ where $p_1 = \nu(1)$ and $p_{k+1} = \nu(p_k + 1) > p_k$. Since $a_{p_{k+1}}$ succeeds a_{p_q} in the sequence it follows by the preceding argument that $a_{p_k} > a_{p_{k+1}}$; that is, $b_k > b_{k+1}$, or $b : k \to b_k$ is decreasing.

Since every sequence has a monotone subsequence, and since every monotone bounded sequence is convergent we see that every bounded sequence has a convergent subsequence. On the other hand any convergent sequence is bounded. For, if $\lim\limits_{k\sim\infty} a_k = \ell$, take $\epsilon = 1$ to obtain

$$|a_k - \ell| < 1 \quad \text{for} \quad k > \omega \quad \text{where} \quad \omega = \Omega(1) .$$

Thus if $M = \max\{|a_1|, |a_2|, \dots, |a_\omega|, |\ell 1 + 1|\}$ we have $|a_k| < M$ for all k and therefore $a : k \to a_k$ is a bounded sequence. Thus boundedness is necessary for convergence and sufficient for the existence of a convergent subsequence. Hence, one way a sequence can fail to converge is that the sequence is not bounded.

If the sequence $a : k \to a_k$ is bounded then we know that it has a convergent subsequence, say $b : k \to a_{i_k}$. Suppose $\lim\limits_{k\sim\infty} a_k = \ell$ and $\lim\limits_{k\sim\infty} a_{i_k} = m$, then m must be ℓ. If $\ell \neq m$, let $\epsilon = \frac{|\ell - m|}{3}$ then for $j > \omega$, $|a_j - \ell| < \epsilon$ and for $k > \omega_2$, $|a_{i_k} - m| < \epsilon$. Thus if $j > \max\{\omega_1, \omega_2\}$ we have

$$|a_j - \ell| < \frac{|\ell - m|}{3} \quad \text{and} \quad |a_j - m| < \frac{|\ell - m|}{3} .$$

Whence $|\ell - m| \leq |\ell - a_j| + |a_j - m| < \frac{2}{3}|\ell - m|$, impossible.

Conversely, suppose the sequence $a : k \to a_k$ is bounded and has the property that all convergent subsequences converge to the same limit, ℓ, then $\lim\limits_{k\sim\infty} a_k = \ell$. Suppose it is false that $\lim\limits_{k\sim\infty} a_k = \ell$, then there exists an $\epsilon > 0$ and a subsequence $b : k \to a_{i_k}$ such that

$$|a_{i_k} - \ell| > \epsilon \quad \text{for} \quad k = 1, 2, \dots .$$

But the subsequence b is also bounded and therefore has a convergent subsequence $c : k \to c_k = b_{j_k} = a_{i_{j_k}}$, which is also a subsequence of a. By Theorem 14-2f, $\lim_{k\sim\infty} c_k > \ell + \epsilon$ or $\lim_{k\sim\infty} c_k < \ell - \epsilon$. In either case $\lim_{k\sim\infty} c_k \neq \ell$ violating the assumption that all convergent subsequencies of a converge to ℓ. Hence $\lim_{k\sim\infty} a_k = \ell$.

The preceding observations can be summarized in the following theorem.

THEOREM 14-2j. For the sequence $a : k \to a_k$ to converge it is both necessary and sufficient that,

1. a is bounded,
2. all convergent subsequences have the same limit.

Cauchy discovered a simple criterion which expresses conditions (1) and (2) simultaneously.

DEFINITION 14-2b. A sequence $a : k \to a_k$ is called a Cauchy sequence if given any $\epsilon > 0$ there exists $\omega = \Omega(\epsilon)$ such that for $k, j > \omega$ we have

$$|a_k - a_j| < \epsilon .$$

THEOREM 14-2k. (Cauchy Convergence Theorem) A sequence $a : k \to a_k$ converges if, and only if it is a Cauchy sequence.

Proof.

(a) If $\lim_{k\sim\infty} a_k = \ell$, then given $\epsilon > 0$ take $\omega = \Omega(\frac{\epsilon}{2})$ so that if $k, j > \omega$ we have $|a_k - \ell| < \frac{\epsilon}{2}$ and $|a_j - \ell| < \frac{\epsilon}{2}$ whence $|a_k - a_j| < \epsilon$.

(b) If $\{a_k\}$ is a Cauchy sequence then,

(i) a is bounded. For the proof of boundedness let $\epsilon = 1$. Then if $k > \nu$ where $\nu = N(1)$, for integral ν, we have $|a_k - a_j| < 1$, with $j = N(1) + 1$; hence a bound for a_k is $M = \max\{|a_1|, |a_2|, \ldots, |a_{j-1}|, |a_j| + 1\}$.

It follows from the boundedness of $\{a_k\}$ that a has a convergent subsequence $k \to a_{i_k}$ with $\lim\limits_{k\to\infty} a_{i_k} = \ell$. We now show

(ii) $\lim\limits_{k\to\infty} a_k = \ell$. For given $\epsilon > 0$ there exists an $\omega_1 = \Omega_1(\epsilon)$ and $\omega_2 = \Omega_2(\epsilon)$ such that for $k > \omega_1$, $|a_{i_k} - \ell| < \epsilon$ and for $k , j > \omega_2$, $|a_k - a_j| < \epsilon$. Thus if $k > \omega$ where $\omega = \max\{\Omega_1(\frac{\epsilon}{2}) , \Omega_2(\frac{\epsilon}{2})\}$ we have

$$|a_k - \ell| \leq |a_k - a_{i_k}| + |a_{i_k} - \ell| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon .$$

Exercises 14-2

1. Prove

(a) Theorem 14-2a
(b) Theorem 14-2b
(c) Theorem 14-2c
(d) Theorem 14-2d
(e) Theorem 14-2e
(f) Theorem 14-2f
(g) Corollary to Theorem 14-2c
(h) Lemma 14-2
(i) Corollary 1 to Lemma 14-2
(j) Corollary 2 to Lemma 14-2
(k) Corollary to Theorem 14-2e
(ℓ) Corollary 1 to Theorem 14-2f
(m) Corollary 2 to Theorem 14-2f

2. Show that if $A_1 < A < A_2$, where $A = \lim\limits_{k\to\infty} a_k$, then there is a number ω such that $k > \omega$ implies $A_1 < a_k < A_2$.

3. Prove that $\lim\limits_{k\to\infty} |a_k| = 0$ if, and only if $\lim\limits_{k\to\infty} a_k = 0$.

4. Let f be a function whose domain contains the point a and points of every deleted neighborhood of a . Prove the converse of Theorem 14-2g. Namely, if $\lim\limits_{n\to\infty} f(x_n) = f(a)$ for every sequence $n \to x_n$ whose terms lie in the domain of f , and which has the limit a , then f is continuous at a .

5. Find $\lim\limits_{n\to\infty} (\sqrt{n^2 + n} - n)$.

6. Find the limits of the following sequences

(a) $n \to (1 + \frac{1}{n^2})^n$;
(b) $n \to \frac{r^n}{n!}$;
(c) $\frac{1}{n^\alpha}$, $\alpha > 0$;
(d) $n \to \frac{\log n}{n^\alpha}$, $\alpha > 0$.

7. Show that $\sqrt{2+\sqrt{2+\sqrt{2+\ldots}}} = 2$; that is, show that the sequence $a : k \to a_k$ defined by $a_1 = \sqrt{2}$, $a_{k+1} = \sqrt{2 + a_k}$ converges and the limit is 2 .

8. Show that $\sqrt{2\sqrt{2\sqrt{2 \ldots}}} = 2$, namely, that the sequence $a : k \to a_k$ defined by $a_1 = \sqrt{2}$, $a_{k+1} = \sqrt{2a_k}$ converges and the limit is 2 .

9. The Fibonacci numbers are defined by $f_0 = f_1 = 1$ and $f_{n+2} = f_n + f_{n+1}$, $n = 0 , 1 , 2 , \ldots$. Find $\lim\limits_{n \to \infty} \frac{f_{n+1}}{f_n}$.

10. Show that the sequence

$$a : n \to a_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} - \log n$$

converges. (The limit of this sequence is called Euler's constant, γ . It is not known whether or not γ is rational.)

11. Given a sequence $a : n \to a_n$, form the sequence

$$\sigma : n \to \sigma_n = \frac{1}{n} \sum_{k=1}^{n} a_k \ .$$

(a) Prove that if $\lim\limits_{n \to \infty} a_n = m$ then $\lim\limits_{n \to \infty} \sigma_n = m$.

(b) Show that σ may converge while a does not.

12. Prove that if $c : k \to c_k$ is a subsequence of $b : k \to b_k$ and b is a subsequence of $a : k \to a_k$, then c is a subsequence of a .

13. Find a sequence with no convergent subsequence.

14. Show that if $k \to a_{i_k}$ is a subsequence of $k \to a_k$ then $i_k \geq k$.

15. Show that if $k \to s_{i_k}$ and $k \to s_{i_k}$ are two subsequences of $n \to s_n$ satisfying $\lim\limits_{k \to \infty} s_{i_k} = \lim\limits_{k \to \infty} s_{j_k} = S$ and the sets of indices i_k and j_k together include all natural numbers, then $\lim\limits_{k \to \infty} s_k = S$.

16. Let $a : n \to a_n$ be a bounded sequence. Let C be the set of limits of subsequences of a . (The elements of C are called cluster points of a .) The least upper bound of C , sup C , is called the limit superior of a and is written $\overline{\lim}\, a_n$. Prove that $\overline{\lim}\, a_n \in C$.

17. As is in Exercise 16, define the limit inferior of a as $\underline{\lim}\, a_n = \inf C$, where inf C is the greatest lower bound or infinum of C . Prove $\underline{\lim}\, a_n \in C$.

18. For each of the following sequences a , find $\overline{\lim}\, a_n$ and $\underline{\lim}\, a_n$.

(a) $a : n \to (-1)^n$

(b) $a : n \to \cos \frac{2n\pi}{5}$

(c) $a : n \to \frac{1}{n}$

(d) $a : n \to \alpha + (1 + (-1)^n)(\frac{\beta - \alpha}{2})$

19. Let $a : n \to a_n$ be bounded , $|a_n| < M$. Suppose that

$$A_1 < \underline{\lim}\, a_n \leq \overline{\lim}\, a_n < A_2 .$$

Prove that there exists an ω such that for $k > \omega$, $A_1 < a_k < A_2$.

20. Suppose that a number A is less than the limit superior of the bounded sequence $a : n \to a_n$, that is, $A < \overline{\lim}\, a_n$. Show that a has a subsequence $b : k \to b_k = a_{i_k}$ satisfying $b_k > A$ for all k .

21. Let f be continuously differentiable and consider the sequences $n \to a_n$ and $n \to b_n$ which both converge to a , where $a_n \neq b_n$ for n = 1 , 2 , ... Show that the sequence

$$n \to \frac{f(a_n) - f(b_n)}{a_n - b_n}$$

converges to f'(a) .

Λ22. Show by an example that the contimity of the derivative is essential in Number 21.

14-3. *Series*.

Since $\lim\limits_{k\to\infty} (\frac{1}{2})^{k+1} = 0$, we know that $\lim\limits_{k\to\infty} \dfrac{1 - (\frac{1}{2})^{k+1}}{1 - \frac{1}{2}} = 2$. But $\dfrac{1 - (\frac{1}{2})^{n+1}}{1 - \frac{1}{2}} = 1 + \frac{1}{2} + (\frac{1}{2})^2 + \ldots + (\frac{1}{2})^n$, thus it is natural to write $2 = 1 + \frac{1}{2} + (\frac{1}{2})^2 + \ldots \sum\limits_{n=0}^{\infty} (\frac{1}{2})^n$. This is an example of the following general situation. We are given a sequence $a : k \to a_k$ and form the sequence $s : k \to s_k = \sum\limits_{j=1}^{k} a_j$; if $\lim\limits_{k\to\infty} s_k = s$ then we write $s = a_1 + a_2 + \ldots = \sum\limits_{i=1}^{\infty} a_i$. The sequence s of *partial sums* s_k is called the *series* associated with the sequence a of terms a_k . Again, we say that the series converges if the sequence of partial sums converge. Thus a series is given by specifying either the sequence of terms a , or the sequence of partial sums s . Given one, the other is defined by one of the relations

$$s_k = a_1 + a_2 + \ldots + a_k \tag{1a}$$

or

$$a_k = s_{k+1} - s_k \, . \tag{1b}$$

Without ambiguity then we shall frequently refer to the series, $\sum\limits_{i=1}^{\infty} a_i$ without regard to the question of its convergence.

It is convenient to consider series written $\sum\limits_{i=k}^{\infty} a_i$ where k is any integer. We are simply considering the sequence of terms $a : i \to a_i$ whose domain is the set of integers n such that $n \geq k$. The sequence of partial sums is defined accordingly by $s_j = \sum\limits_{i=k}^{k+j-1} a_i$. Most frequently k will be either 0 or 1 .

THEOREM 14-3a. If $\sum_{i=1}^{\infty} a_i$ and $\sum_{i=1}^{\infty} b_i$ converge and λ is a real number, then $\sum_{i=1}^{\infty} \lambda a_i$ converges to $\lambda \sum_{i=1}^{\infty} a_i$ and $\sum_{i=1}^{\infty} (a_i + b_i)$ converges to $\sum_{i=1}^{\infty} a_i + \sum_{i=1}^{\infty} b_i$.

Proof. Use Theorems 14-2b and c.

Corollary. A linear combination of series converges to the same linear combination of their sums.

We state several useful criteria for convergence and divergence.

THEOREM 14-3b. (Cauchy Criterion) The series $\sum_{i=1}^{\infty} a_i$ converges if, and only if to every $\epsilon > 0$ there exists $\omega = \Omega(\epsilon)$ such that if $m > n > \omega$ then

$$\left|\sum_{i=n}^{m} a_i\right| < \epsilon .$$

Proof. Observe that $|s_m - s_{n-1}| = \left|\sum_{i=n}^{m} a_i\right|$ and apply the Cauchy Convergence Theorem.

Corollary. (n-th term test) If the series $\sum_{i=1}^{\infty} a_i$ converges then $\lim_{i \to \infty} |a_i| = 0$.

Example 14-3a. The series $\sum_{n=1}^{\infty} (-1)^n$ does not converge by the n-th term test. Note however that $\sum_{n=0}^{\infty} [(-1)^{2n+1} + (-1)^{2n+2}] = 0$ while $-1 + \sum_{n=1}^{\infty} [(-1)^{2n} + (-1)^{2n+1}] = -1$. Euler used this observation to conclude that $\sum_{n=1}^{\infty} (-1)^n = -\frac{1}{2}$! As ludicrous as Euler's conclusion may seem this makes some sense in the light of Exercises 14-2, Number 2.

While the n-th term test is a necessary condition for convergence it is not sufficient. The following is an example of the insufficiency of the n-th term test.

Example 14-3b. The series $\sum_{n=1}^{\infty} \frac{1}{n}$ does not converge. I.e., the series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges. For proof, we group terms, and observe that

$$1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{2^n - 1} + \frac{1}{2^n} = 1 + \frac{1}{2} + (\frac{1}{3} + \frac{1}{4}) + (\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}) + \dots .$$

In general, $\sum_{n=2^k+1}^{2^{k+1}} \frac{1}{n} \geq \sum_{n=2^k+1}^{2^{k+1}} \frac{1}{2^{k+1}} = \frac{1}{2^{k+1}} \sum_{n=2^k+1}^{2^{k+1}} 1 = \frac{2^k}{2^{k+1}} = \frac{1}{2}$; but given any ω there is a k such that $2^k > \omega$. Thus with $n = 2^k$, $m = 2^{k+1}$ and $\epsilon < \frac{1}{2}$, Theorem 14-3b shows that the series diverges.

THEOREM 14-3c. (First Comparison Test) If $0 \leq a_k \leq b_k$ for $k = 1 , 2 , \dots ,$ and the series $\sum_{i=1}^{\infty} b_i$ converges then the series $\sum_{i=1}^{\infty} a_i$ converges and $\sum_{i=1}^{\infty} a_i \leq \sum_{i=1}^{\infty} b_i$.

Proof. Let $s : k \to s_k = \sum_{i=1}^{k} a_i$ and $\tau : k \to \tau_k = \sum_{i=1}^{k} b_i$, then $s \leq \tau$. Since $a_i \geq 0$ the sequence s is monotone increasing. Since τ converges it is bounded. Thus s is also bounded and therefore convergent. That $\sum_{i=1}^{\infty} a_i \leq \sum_{i=1}^{\infty} b_i$, follows from the Sandwich Theorem.

Corollary. If $0 \leq a_k \leq b_k$ for $k = 1 , 2 , \ldots$, and the series $\sum_{i=1}^{\infty} a_i$ diverges then the series $\sum_{i=1}^{\infty} b_i$ diverges.

In both the theorem and the corollary we can replace the hypothesis by $0 \leq a_k \leq b_k$ for $k > \omega$.

THEOREM 14-3d. (Integral Test) If

(1) $f : \tau \to f(\tau)$ is a monotone decreasing continuous function.

(2) $f(k) = a_k$, $k = 1 , 2 , \ldots$.

Then the integral $\int f(\tau)d\tau$ and the series $\sum_{k=1}^{\infty}$ either both converge or both diverge.

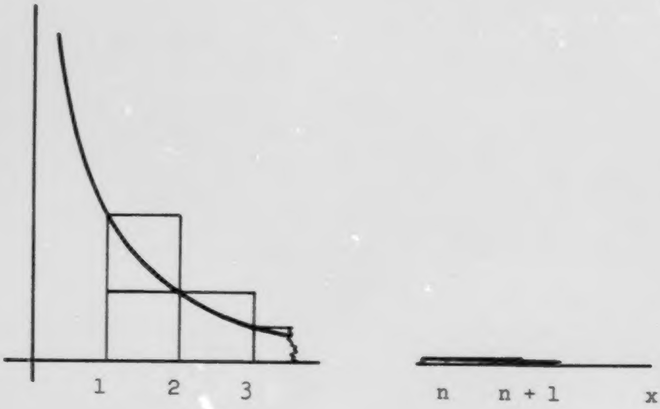


Figure 14-3a

Proof. We have

$$\sum_{i=1}^{n} a_i \leq a_1 + \int_1^n f(\tau)d\tau \quad \text{and} \quad \sum_{i=1}^{n} a_i \geq \int_1^{n+1} f(\tau)d\tau \ .$$

(See Figure 14-3a.) Thus the conclusion follows from the Monotone Convergence Theorem (Theorem 14-2h), since $a_i > 0$ for all i and f is positive.

THEOREM 14-3e. (p - test) The series $\sum_{n=1}^{\infty} \frac{1}{n} p$ converges if $p > 1$ and diverges of $p \leq 1$.

Proof. Use the Integral Test and Theorem 10-6b.

Example 14-3c. The series $\sum_{n=1}^{\infty} \frac{\sqrt{n}}{n^2 + 1}$ converges. For $\frac{\sqrt{n}}{n^2 + 1} < \frac{\sqrt{n}}{n^2} = \frac{1}{n^{3/2}}$ and $\sum_{n=1}^{\infty} \frac{1}{n}^{3/2}$ converges by the p-test.

Example 14-3d. The series $\sum_{n=2}^{\infty} \frac{1}{n \log n}$ diverges but $\sum_{n=2}^{\infty} \frac{1}{n(\log n)^2}$ converges, since $\int_2^{\infty} \frac{d\tau}{\tau \log \tau}$ diverges while $\int_2^{\infty} \frac{d\tau}{\tau(\log \tau)^2}$ converges.

The First Comparison Test is sometimes inconvenient to apply. For example, if we are looking at the series $\sum_{n=1}^{\infty} \frac{\sqrt{\sqrt{6n^3} + \sqrt{n^2 + 1}}}{n^3 + 2n - 1}$, we feel that this series is "essentially like" the series $\sum_{n=1}^{\infty} \frac{\sqrt{6}}{n^{3/2}}$. However, it is a tedious job to derive the inequalities needed for the First Comparison Test. The following theorem makes precise and justifies the phrase "essentially like."

THEOREM 14-3f. (Second Comparison Test) If $\lim\limits_{k\to\infty} \frac{a_k}{b_k} = c \neq 0$ then the series $\sum\limits_{n=1}^{\infty} a_k$ and $\sum\limits_{k=1}^{\infty} b_k$ either both converge or both diverge.

Proof. If $c > 0$, then by Theorem 14-2 there exists an ω such that if $k > \omega$ then

$$\frac{c}{2} < \frac{a_k}{b_k} < \frac{3c}{2} .$$

Thus $\frac{c}{2} b_k < a_k < \frac{3}{2} c\, b_k$, for $k > \omega$. Now apply Theorem 14-3a and the First Comparison Test.

Example 14-3e. The series $\sum\limits_{n=1}^{\infty} \frac{\sqrt{6n^3 + \sqrt{n^2 + 1}}}{n^3 + 2n - 1}$ converges, since

$$\lim_{n\to\infty} \frac{\sqrt{6n^3 + \sqrt{n^2 + 1}}}{n^3 + 2n - 1} \Bigg/ \frac{1}{n^{3/2}} = \sqrt{6} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{1}{n^{3/2}} \text{ converges by the p-test.}$$

At the beginning of this section we gave a special case of the geometric series $\sum\limits_{n=0}^{\infty} r^n$. If $|r| < 1$ then $\sum\limits_{n=0}^{\infty} r^n \frac{1}{1 - r}$ since $\lim\limits_{n\to\infty} |r|^n = 0$.

The next two theorems are tests for convergence which exploit our knowledge of the convergence of the geometric series $\sum\limits_{n=0}^{\infty} r^n$. What is needed is a simple condition on a positive sequence $a : k \to a_k$ to ensure that $a_k < c\, r^k$ for some $r < 1$ and $k > \omega$.

THEOREM 14-3g. (Ratio Test) Let $a : k \to a_k$ be a positive sequence for which $\lim\limits_{k\to\infty} \frac{a_{k+1}}{a_k} = r$. Then the series $\sum\limits_{k=0}^{\infty} a_k$ converges if $r < 1$ and diverges if $r > 1$.

Proof. The proof of Corollary 1 of Lemma 14-2 can be extended (Exercises 14-2, No. 2) to prove that if $\lim\limits_{k\sim\infty} a_k = c$ and $c_1 < c < c_2$ then there exists ω such that for $k > \omega$.

$$c_1 < a_k < c_2 \ .$$

If $r < 1$, choose r , so that $r < r_1 < 1$. Then there exists an integer ν , such that for $k > \nu$,

$$\frac{a_{k+1}}{a_k} < r_1$$

or

$$a_{k+1} < r_1 a_k \ .$$

Then by induction we have $a_{\nu+p} < r_1{}^p a_\nu$ and, therefore, $\sum\limits_{k=0}^{\infty} a_k$ converges by comparison with $\sum\limits_{k=0}^{\infty} \frac{r_1{}^k a_\nu}{r_1{}^\nu}$.

On the other hand, if $r > 1$ then choose r_2 so that $1 < r_2 < r$, where there is an integer ν such that for $k > \nu$

$$\frac{a_{k+1}}{a_k} > r_2 > 1 \ .$$

Then, for $k > \nu$, $a_{k+1} > a_k$ and therefore $a_k > a_{\nu+1}$. Hence, by the Sandwich Theorem we cannot have $\lim\limits_{k\sim\infty} a_k = 0$, and by the n-th term test $\sum\limits_{k=0}^{\infty} a_k$ diverges.

THEOREM 14-3h. (Root Test) Let $a : k \to a_k$ be a positive sequence for which

$\lim\limits_{n\sim\infty} \sqrt[n]{a_n} = r$, then $\sum\limits_{n=1}^{\infty} a_n$ converges if $r < 1$ and diverges if $r > 1$.

Proof. If $r < 1$, choose r_1 so that $r < r_1 < 1$ and ω such that, for $k > \omega$,

$$\sqrt[k]{a_k} < r_1 \text{ or } a_k < r_1^{\,k} ,$$

and, therefore, $\sum_{k=0}^{\infty} a_k$ converges by comparison with $\sum_{k=0}^{\infty} r_1^{\,k}$. On the other hand if $r > 1$, choose r_2 so that $1 < r_2 < r$ and ω such that, for $k > \omega$, $a_k > r_1$ or $a_k > r_2^{\,k} > 1$. Thus we cannot have $\lim_{k\sim\infty} a_k = 0$ and $\sum_{k=0}^{\infty} a_k$ diverges.

Example 14-3f. The series $\sum_{n=1}^{\infty} \frac{n^2 + 1}{n!}$ converges. Since

$$\lim_{n\sim\infty} \frac{(n+1)^2 + 1}{(n+1)!} \Big/ \frac{n^2 + 1}{n!} = \lim_{n\sim\infty} \frac{1}{n+1} \, \frac{(n+1)^2 + 1}{n^2 + 1} = \lim_{n\sim\infty} \frac{1}{n+1} = 0 .$$

Exercises 14-3

1. Test the following series for convergence.

(a) $\sum_{n=1}^{\infty} \frac{\sqrt{n+1} - \sqrt{n}}{n}$

(b) $\sum_{n=3}^{\infty} \frac{1}{n(\log n)(\log \log n)^2}$

(c) $\sum_{n=1}^{\infty} \frac{(n+1)2^n}{n3^n}$

(d) $\sum_{n=1}^{\infty} \frac{n^3}{n!}$

2. Does the series $\sum_{n=1}^{\infty} (\frac{n}{n+1})^{n^2}$ converge?

3. Find a suitable $\omega = \Omega(\epsilon)$ for each of the following series.

   (a) $\sum_{n=2}^{\infty} \frac{1}{n \log^2 n}$

   (b) $\sum_{n=0}^{\infty} \frac{n}{3^n}$

   (c) $\sum_{n=1}^{\infty} \frac{\sqrt{n - 1}}{n^2}$

4. Show that if $a_n \geq 0$, $n = 1 , 2 , \ldots$, and $\sum_{n=1}^{\infty} a_n$ converges then

   $\sum_{n=1}^{\infty} \frac{\sqrt{a_n}}{n}$ converges.

5. Let $a : k \to a_k$ be a monotone decreasing sequence. Show that if a has a subsequence $k \to a_{i_k}$ for which $a_{i_k} > \frac{1}{i_k}$ then $\sum_{k=1}^{\infty} a_k$ diverges.

6. Show that if the series of positive terms $\sum_{i=1}^{\infty} a_i$ diverges then

   $\sum_{i=1}^{\infty} \frac{a_i}{s_i}$ diverges, where $s_i = \sum_{k=1}^{i} a_k$.

7. Show that if the series of positive terms $\sum_{n=1}^{\infty} a_n$ converges then

   $\sum_{n=1}^{\infty} {a_n}^2$ converges.

8. Prove that $\sum_{n=1}^{\infty} \frac{(-1)^n}{n} = \log 2$.

   Hint: Use $\sum_{k=0}^{n} x^k + \frac{x^{n+1}}{1 - x} = \frac{1}{1 - x}$ and integrate.

9. (Cauchy Condensation Test) Show that if $n \to a_n$ is a decreasing sequence of positive terms then the series $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=0}^{\infty} 2^n a_{2^n}$ either both converge or both diverge.

10. (a) Use the Cauchy Condensation Test to show that $\sum_{k=2}^{\infty} \frac{1}{n \log n}$ diverges and that $\sum_{n=2}^{\infty} \frac{1}{n(\log n)^2}$ converges.

 (b) Apply the Cauchy Condensation Test to test the convergence of $\sum_{n=2}^{\infty} \frac{1}{n(\log n)(\log \log n)}$.

14-4. Conditional and Absolute Convergence.

So far we have considered primarily series with positive terms. The sequence of partial sums has generally been monotone and we have given conditions which guaranteed that the terms a_k became small "fast enough." However, if the series contains both positive and negative terms then cancellation effects might contribute to the convergence of the series.

Thus, as with improper integrals, if $\sum_{i=1}^{\infty} |a_i|$ converges we say that the series $\sum_{i=1}^{\infty}$ *converges absolutely* while if $\sum_{i=1}^{\infty} a_i$ converges and $\sum_{i=1}^{\infty} |a_i|$ diverges we say that $\sum_{i=1}^{\infty}$ *converges conditionally*.

To justify the phrase "absolute convergence" we prove that absolute convergence: implies convergence.

THEOREM 14-4a. If $\sum_{i=1}^{\infty} |a_i|$ converges then $\sum_{i=1}^{\infty} a_i$ converges.

Proof. Let $s : k \to s_k = \sum_{i=1}^{k} a_i$ and $t : k \to t_k = \sum_{i=1}^{k} |a_i|$. Then t is a Cauchy Sequence, and for $n > m$

$$|s_n - s_m| = \left| \sum_{i=m+1}^{n} a_i \right| < \left| \sum_{i=m+1}^{n} |a_i| \right| = |t_n - t_m| .$$

Thus s is also a Cauchy Sequence and therefore converges.

With Theorem 14-4a all the tests for convergence of Section 14-3 can be applied to series with positive and negative terms by testing the given series for absolute convergence.

Example 14-4a. The series $\sum_{n=1}^{\infty} \frac{(-1)^n}{n^2}$ converges since the series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges by the p-test.

We have exhibited many tests for absolute convergence. The following test is the principal means for establishing the convergence of conditionally convergent series.

THEOREM 14-4b. (Leibniz's Test for Alternating Series) If $a : k \to a_k$ is such that

1. $\operatorname{sgn} a_{k+1} = -\operatorname{sgn} a_k$ for $k = 1 , 2 , \ldots ,$
2. $|a| : k \to |a_k|$ is monotone decreasing,
3. $\lim_{k \to \infty} |a_k| = 0$,

then $\sum_{k=1}^{\infty} a_k$ converges.

Condition 1 states that the terms alternate in sign. Such a series is called an alternating series.

Proof. Under Condition 2, the sums of two consecutive terms has the same sign as the first of the pair:

$$\operatorname{sgn}(a_k + a_{k+1}) = \operatorname{sgn} a_k .$$

Now consider the case $a_1 > 0$. (If $a_1 < 0$, the proof may be applied directly to the series $-\sum_{k=1}^{\infty} a_k$.) The sum of an even number of terms can be obtained by associating sums of consecutive pairs as follows,

$$S_{2n} = (a_1 + a_2) + (a_3 + a_4) + \ldots + (a_{2n-1} + a_{2n}) .$$

Since $a_{2k-1} + a_{2k} > 0$, it follows that S_{2n} defines an increasing sequence $s_e : n \to S_{2n}$. Similarly, since $a_{2k} + a_{2k+1} < 0$, the sums of odd numbers of terms,

$$S_{2n+1} = a_1 + (a_2 + a_3) + \ldots + (a_{2n} + a_{2n+1})$$

form a decreasing sequence $s_0 : n \to S_{2n+1}$. The sequences s_e and s_0 are not only monotone, but also bounded, since

$$S_{2n} = S_{2n-1} + a_{2n} < S_{2n-1} \leq S_1 \leq a_1$$

and

$$S_{2n+1} = S_{2n} + a_{2n+1} > S_{2n} \geq S_2 \geq a_1 + a_2 ;$$

therefore, by the Monotone Convergence Theorem, both s_e and s_0 converge. Moreover, since the sequence $n \to S_{2n+1} - S_{2n} = a_{n+1}$ converges to 0 , both sequences s_0 and s_e have the same limit, S . It follows directly that $\sum_{k=1}^{\infty} a_k$ also has the sum S (Exercises 14-2, No. 15).

Finally, note that we have a simple estimate of the error of approximation of the n-th partial sum S_n to the limit S : since $S_{2i} < S < S_{2j+1}$ for any i and j (Corollary to Theorem 14-2h) it follows for all n that

$$|S - S_n| < |a_{n+1}| .$$

In other words, the absolute error in approximating the limit of a convergent alternating series by the sum to n terms is less than the magnitude of the next term.

Example 14-4b. The series $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ converges conditionally since $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges while $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ converges since the hypotheses of Theorem 14-4b are satisfied.

Exercises 14-4

1. Show that if $\sum_{n=1}^{\infty} a_n^2$ and $\sum_{n=1}^{\infty} b_n^2$ both converge, then $\sum_{n=1}^{\infty} a_n b_n$ converges.

2. Test $\sum_{n=1}^{\infty} (-1)^n \frac{n^2}{(n+1)^2}$ for convergence.

3. Is the following true in general? If $\sum_{n=1}^{\infty} a_n$ converges and $\lim_{n\to\infty} c_n = 0$, then $\sum_{n=1}^{\infty} a_n c_n$ converges.

4. Test for convergence the alternating series $\sum_{k=1}^{\infty} a_k$, where $a_{2k+1} = \frac{1}{k}$ and $a_{2k} = -\frac{1}{2^k}$.

5. Prove that if $|a_{n+1}a_{n-1}| < a_n^2$ for all n and $|a_2| < |a_1|$, then $\sum_{n=1}^{\infty} a_n$ converges absolutely.

14-5. Parentheses and Rearrangements.

From Example 14-3a we see that the insertion of parentheses in a divergent series can make a new series which converges. Consider what happens if we insert parentheses in a convergent series. Let $\sum_{i=1}^{\infty} a_i = \ell$ and $s : k \to s_k = \sum_{i=1}^{k} a_i$ and let $\sum_{i=1}^{\infty} b_i$ be a series obtained from $\sum_{i=1}^{\infty} a_i$ by the insertion of parentheses. A typical term of the sequence $b : k \to b_k$ has the form

$$b_i = a_{j_i} + a_{j_i+1} + \dots + a_{j_i+k_i} ,$$

where

$$j_{i+1} = j_i + k_i + 1 .$$

Therefore, the sequence of partial sums, $t : k \to t_k = \sum_{i=1}^{k} b_i$, is a subsequence of s , $t_n = s_{(j_n+k_n)}$, and thus $\lim_{k\to\infty} t_k = \lim_{k\to\infty} s_k = \ell$. So, insertion of parentheses in a convergent sequence doesn't affect its sum.

For a finite sum $a_1 + a_2 + \dots + a_n$ a rearrangement of the summands does not affect the sum. The situation is not quite as simple for series. The following two theorems tell us when we can rearrange.

THEOREM 14-5a. If the series $\sum_{i=1}^{\infty} a_i$ converges to S absolutely, then any rearrangement of the series also converges to S .

Proof. Let $\sum_{i=1}^{\infty} b_i$ be a rearrangement of $\sum_{i=1}^{\infty} a_i$. We outline the idea of the proof Choose a ν so large that the partial sums $\sum_{i=1}^{\nu} |a_i|$, and hence $\sum_{i=1}^{\nu} a_i$, closely approximate the sums of the corresponding infinite series. For the rearrangement b of the series, take a partial sum which includes all the terms $a_1, a_2, \ldots, a_\nu$. The remaining terms of a, no matter how distributed in the rearrangement, can have only a slight effect; therefore, the partial sum closely approximates $\sum_{i=1}^{\nu} a_i$ and, hence, $\sum_{i=1}^{\infty} a_i$. The proof consists of a careful accounting for the small effect of the terms beyond a_ν.

Let $\sum_{i=1}^{\infty} |a_i| = S^*$, $\sum_{i=1}^{\infty} a_i = S$, $s : k \to s_k = \sum_{i=1}^{k} a_i$, $t : k \to t_k = \sum_{i=1}^{k} b_i$, and $s^* : k \to s_k^* = \sum_{i=1}^{k} |a_i|$. Given $\epsilon > 0$, there exists an integer $\nu = N(\epsilon)$ such that if $k > \nu$ then $|S^* - s_k^*| = \sum_{i=k+1}^{\infty} |a_i| < \epsilon$ and $|S - s_k| = |\sum_{i=k+1}^{\infty} a_i| < \epsilon$. That $\sum_{i=1}^{\infty} b_i$ is a rearrangement of $\sum_{i=1}^{\infty} a_i$ means that we have a one-to-one mapping $\psi : i \to \psi(i)$ of the set of natural numbers onto itself and $b_i = a_{\psi(i)}$. We may also consider $\sum_{i=1}^{\infty} a_i$ as a rearrangement $\sum_{i=1}^{\infty} b_i$ by introducing the inverse mapping to ψ, that is, $\phi : \psi(i) \to i$. Thus $b_j = a_{\phi(j)}$. Set $\nu_1 = \max\{\phi(j) : 1 \le j \le \nu\}$. Then, if $k > \nu_1$, we have

$$\begin{aligned}
|S - t_k| &= |S - \sum_{j=1}^{k} b_j| \\
&= |S - \sum_{i=1}^{\nu} b_{\psi(i)} - \sum_{\substack{\phi(j)>\nu \\ j \le k}} b_j| \\
&= |S - \sum_{i=1}^{\nu} a_i - \sum_{\substack{\phi(j)>\nu \\ j \le k}} a_{\phi(j)}| \\
&\le |S - \sum_{i=1}^{\nu} a_i| + \sum_{\substack{\phi(j)>\nu \\ j \le k}} |a_{\phi(j)}| \quad \text{(formula continued)}
\end{aligned}$$

$$\leq \left|S - \sum_{i=1}^{\nu} a_i\right| + \sum_{\phi(j)>\nu} |a_{\phi(j)}|$$

$$\leq \left|S - \sum_{i=1}^{\nu} a_i\right| + \sum_{i=\nu+1}^{\infty} |a_i|$$

$$\leq \left|S - \sum_{i=1}^{\nu} a_i\right| + \left|S^* - \sum_{i=1}^{\nu} |a_i|\right|$$

$$< 2\epsilon \ .$$

Thus t_k has S as its limit, which proves the theorem.

THEOREM 14-5b. If $\sum_{i=1}^{\infty} a_i$ is conditionally convergent, then given any number r there is a rearrangement of $\sum_{i=1}^{\infty} a_i$ whose sum is r.

Lemma 14-5. Let $\sum_{i=1}^{\infty} a_i$ be conditionally convergent. Take

$$a^+ : k \to a_k^+ = \begin{cases} a_k \ , \text{ for } a_k \geq 0 \\ 0 \ , \text{ for } a_k < 0 \end{cases}$$

and

$$a^- : k \to a_k^- = \begin{cases} a_k \ , \text{ for } a_k \leq 0 \\ 0 \ , \text{ for } a_k > 0 \ . \end{cases}$$

Then both $\sum_{k=1}^{\infty} a_k^+$ and $\sum_{k=1}^{\infty} a_k^-$ diverge monotonically.

Proof. If $\sum_{k=1}^{\infty} a_k^+$ converged then $\sum_{k=1}^{\infty} a_k^- = \sum_{k=1}^{\infty} a_k^+ - \sum_{k=1}^{\infty} a_k^+$ would also converge and, therefore, $\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\infty} a_k^+ + \sum_{k=1}^{\infty} a_k^-$ being the sum of two convergent series would converge.

Proof of Theorem 14-5b. Consider the case $r > 0$. By Lemma 14-5, we may choose i_1 so that $\sum_{k=1}^{i_1} a_k^+ < r \leq \sum_{k=1}^{i_1+1} a_k^+$. Choose $i_2 > i_1 + 1$ so that

$$\sum_{k=1}^{i_1+1} a_k^+ + \sum_{k=i_1+2}^{i_2} a_k^- > r > \sum_{k=1}^{i_1+1} a_k^+ + \sum_{k=i_1+2}^{i_2+1} a_k^- .$$

Continuing in this manner we construct a rearrangement which converges to r since $\lim_{k\to\infty} |a_i| = 0$. The proofs for the case $r = 0$ and $r < 0$ are similar. (See Exercises 14-5, No. 1.)

We apply Theorem 14-5a to prove the following theorem.

THEOREM 14-5c. If the series $\sum_{i=0}^{\infty} a_i$ and $\sum_{j=0}^{\infty} b_j$ are absolutely convergent to A and B , respectively, then their Cauchy Product $\sum_{i=0}^{\infty} c_i$ converges to AB , where $c_n = \sum_{k=0}^{n} a_k b_{n-k}$.

Proof. We begin with an intuitive approach. Let us expand $\left(\sum_{i=0}^{\infty} a_i\right)\left(\sum_{j=0}^{\infty} b_j\right)$ as by the distributive law*. We obtain the sum of all terms of the form $a_i b_j$ for $i = 0 , 1 , 2 , \ldots , j = 0 , 1 , 2 , \ldots$. Arrange these terms in an array as in Figure 14-5a.

---

*The distributive law holds for finite series. Its use for infinite series has to be explained and justified when it is applicable. Here we only use it to suggest the scheme of the proof.

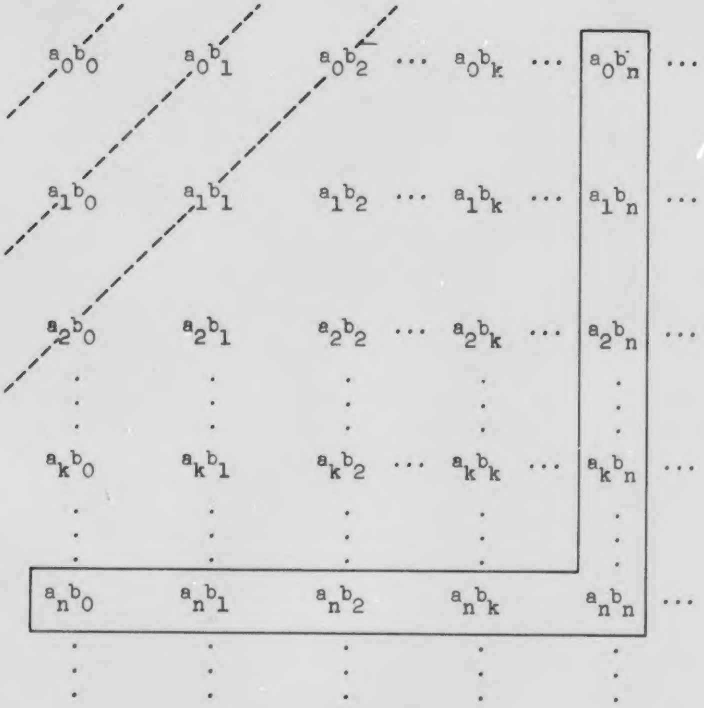

a_0b_0 a_0b_1 a_0b_2 ... a_0b_k ... a_0b_n ...
a_1b_0 a_1b_1 a_1b_2 ... a_1b_k ... a_1b_n ...
a_2b_0 a_2b_1 a_2b_2 ... a_2b_k ... a_2b_n ...
a_kb_0 a_kb_1 a_kb_2 ... a_kb_k ... a_kb_n ...
a_nb_0 a_nb_1 a_nb_2 a_nb_k a_nb_n ...

Figure 14-5a

We see that c_n is the sum of the terms in the n-th diagonal of the array, and

$$e_n = a_nb_0 + a_nb_1 + \dots + a_nb_k + \dots + a_nb_n + \dots + a_{n-1}b_n + \dots$$
$$+ a_kb_n + \dots + a_1b_n + a_0b_n$$

$$= \sum_{\max(i,j)=n} a_ib_j \ ;$$

This is the sum of the terms in the n-th "ell" (called a gnomon by the Greeks). Note that

$$E_{n+1} = \sum_{k=0}^{n} e_k = A_{n+1}B_{n+1} = \left(\sum_{k=0}^{n} a_k\right)\left(\sum_{k=0}^{n} b_k\right).$$

Hence, by Theorem 14-2d, the sequence $n \to E_n$ converges to AB . But

$$|C_n - E_n| = \Big|\sum_{\substack{i+j>n \\ i,j\le n}} a_i b_j\Big| \le \sum_{\substack{i+j>n \\ i,j\le n}} |a_i||b_j|$$

$$\le \left(\sum_{i=0}^{n} |a_i|\right)\left(\sum_{\frac{n}{2}<j\le n} |b_j|\right) + \left(\sum_{\frac{n}{2}<i\le n} |a_i|\right)\left(\sum_{j=0}^{n} |b_j|\right)$$

$$\le \left(\sum_{i=0}^{\infty} |a_i|\right)\left(\sum_{\frac{n}{2}<j} |b_j|\right) + \left(\sum_{\frac{n}{2}<i} |a_i|\right)\left(\sum_{j=0}^{\infty} |b_j|\right)$$

$$\le A^*(B^* - B_\nu^*) + (A^* - A_\nu^*)B^* ,$$

where $A_n^* = \sum_{i=0}^{n-1} |a_i|$, $B_n^* = \sum_{j=0}^{n-1} |b_j|$, $A^* = \sum_{i=0}^{\infty} |a_i|$, $B^* = \sum_{j=0}^{\infty} |b_j|$, and $\nu = \left[\frac{n}{2}\right] + 1$. Since the last sum converges to 0 ,

$$\sum_{i=0}^{\infty} C_i = \lim_{n\sim\infty} C_n = \lim_{n\sim\infty} E_n = \lim_{n\sim\infty} A_n B_n = AB .$$

Exercises 14-5

1. Prove Theorem 14-5b for the case $r < 0$.

2. Show that $\left(\sum_{n=0}^{\infty} \frac{x^n}{n!}\right)\left(\sum_{n=0}^{\infty} \frac{y^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{(x+y)^n}{n!}$ to conclude that $e^x e^y = e^{x+y}$. (See Section 8-5.)

14-6. Sequences of Functions, Uniform Convergence.

In Section 14-3 when we observed that for $|r| < 1$ the geometric series $\sum_{n=0}^{\infty} r^n$ converges to $\frac{1}{1-r}$ we were thinking of an infinite class of series, one series for each r, $|r| < 1$. However, we can also think of it as a series of functions. Namely, we have $u : n \to u_n$ where u_n is the function $u_n : r \to r^n$ and $s : n \to s_n$ where the function s_n is $s_n : r \to \sum_{k=0}^{n} r^k$.

Then the series of functions $\sum_{n=0}^{\infty} u_n$ converges to the function $f : r \to \frac{1}{1-r}$ in the sense, that for each r, $|r| < 1$, the series $\sum_{n=0}^{\infty} u_n(r)$ converges $f(r)$.

DEFINITION 14-6a. The sequence of functions $u : n \to u_n$ where u_n, $x \to u_n(x)$ (let E be the domain of u_n for $n = 1, 2, \ldots$) converges pointwise to the function $f : x \to f(x)$ if for every $x \in E$ and every $\epsilon > 0$ there exists $\omega = \Omega(x, \epsilon)$ such that for $k > \omega$, we have

$$|u_k(x) - f(x)| < \epsilon .$$

Since we have defined a series as a sequence of partial sums, Definition 14-6a immediately defines pointwise convergence for a series. In general, any statement about sequences is directly translatable in this way (i.e., by Equation (1a) of Section 14-3) as a statement about series. We make no use of the fact here, but any statement about series is also directly translatable as a statement about sequences (by Equation (1b) of Section 14-3).

Pointwise convergence is not the "right" idea of convergence for application of the operations of the calculus to a sequence, as the following examples show.

Example 14-6a. Consider the sequence of functions

$$u_n : x \to \begin{cases} 0 , & \text{for } \frac{1}{2^n} \le x \le 1 \\ 1 - 2^n x , & \text{for } 0 \le x \le \frac{1}{2^n} \end{cases}$$

(see Figure 14-6a).

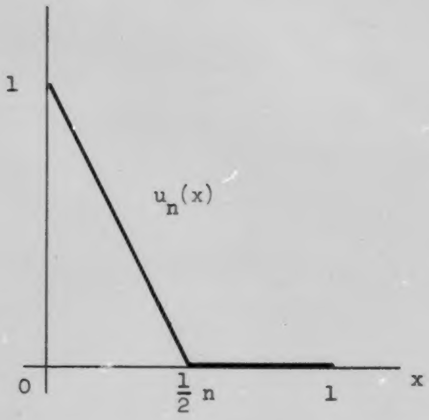


Figure 14-6a

Even though u_n is continuous for $n = 1 , 2 , \ldots$, we have

$$\lim_{n\to\infty} u_n = f : x \to \begin{cases} 1 , & \text{for } x = 0 \\ 0 , & \text{for } 0 < x \le 1 , \end{cases}$$

so that f is discontinuous at 1 . Thus, for pointwise convergence, the limit of a sequence of continuous functions can be discontinuous.

Example 14-6b. Consider the sequence of functions

$$u_n : x \to \begin{cases} 4n^2 x , & \text{for } 0 \le x \le \frac{1}{2n} \\ -(x - \frac{1}{n})4n^2 , & \text{for } \frac{1}{2n} \le x \le \frac{1}{n} \\ 0 , & \text{for } \frac{1}{n} \le x \le 1 \end{cases}$$

(see Figure 14-6b).

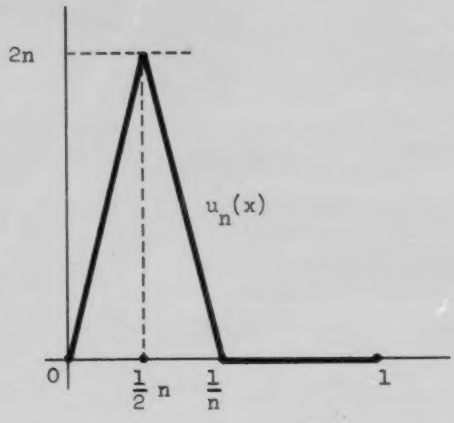


Figure 14-6b

Then $\lim_{n\sim\infty} u_n = 0$ while $\lim_{n\sim\infty} \int_0^1 u_n(x)dx = 1$. Thus

$$\int_0^1 \left[\lim_{n\sim\infty} u_n(x)\right] dx \neq \lim_{n\sim\infty} \int_0^1 u_n(x)dx .$$

Thus for pointwise convergence the integral of the limit might not be the limit of the integrals.

We should very much like to have the properties which Example 14-6a and Example 14-6b show we do not have for pointwise convergence of functions.

The trouble with Definition 14-6a is that ω , the number of terms of the sequence which we must take to be within ϵ of the limit, depends on the particular value of x with which we are concerned. That is, for ϵ fixed the function Ω given by $\omega = \Omega(x,\epsilon)$ is not necessarily bounded. If it is bounded we can choose $\omega = \sup\{\Omega(x,\epsilon) : x \text{ in } E\}$, which is independent of x . Notice that in Example 14-6a and Example 14-6b it is the case that $\Omega(x,\epsilon)$ cannot be bounded and therefore we cannot choose ω independently of x (Exercises 14-6, No. 4).

DEFINITION 14-6b. Consider the sequence of functions $u : n \to u_n$ where $u_n : x \to u_n(x)$ and let E be the domain of u_n , for $n = 1 , 2 , \ldots$. The sequence u *converges uniformly* on E to the function $f : x \to f(x)$ if for every $\epsilon > 0$ there exists $\omega = \Omega(\epsilon)$ such that for $k > \omega$, we have $|u_k(x) - f(x)| < \epsilon$ for all x in E .

Now we shall show that with uniform convergence the phenomena exhibited in Example 14-6a and Example 14-6b are impossible.

THEOREM 14-6a. (The uniform limit of a sequence of continuous function is continuous.) If the sequence u converges uniformly to f on E, and u_n is continuous on E for $n = 1, 2, \dots$, then f is continuous on E.

Proof. Take $x_0 \in E$ and $\epsilon > 0$. We want to estimate $|f(x) - f(x_0)|$ by bounding $|x - x_0|$. Now

$$\begin{aligned} |f(x) - f(x_0)| &= |f(x) - u_n(x) + u_n(x) - u_n(x_0) + u_n(x_0) - f(x_0)| \\ &\leq |f(x) - u_n(x)| + |u_n(x) - u_n(x_0)| + |u_n(x_0) - f(x_0)| . \end{aligned}$$

Take n fixed $> \omega(\frac{\epsilon}{3})$, choose δ so that

$$|u_n(x) - u_n(x_0)| < \frac{\epsilon}{3} \text{ for } |x - x_0| < \delta .$$

Then

$$|f(x) - f(x_0)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \leq \epsilon , \text{ for } |x - x_0| < \delta .$$

THEOREM 14-6b. (The integral of the uniform limit is the limit of the integrals.) If the sequence $u : n \to u_n$ of continuous functions u_n converges uniformly to f on E, and if $[a,b]$ is contained in E then

$$\lim_{n\to\infty} \int_a^b u_n(x)dx = \int_a^b f(x)dx = \int_a^b \left[\lim_{n\to\infty} u_n(x)\right] dx .$$

Proof. By Theorem 14-6a f is continuous and therefore integrable. Given $\epsilon > 0$, there exists $\omega = \Omega(\epsilon)$ such that if $k > \omega$ then $|f(x) - u_k(x)| < \epsilon$ for all x in E. Hence if $k > \Omega(\frac{\epsilon}{b - a})$ then

$$\begin{aligned} \left|\int_a^b f(x)dx - \int_a^b u_k(x)dx\right| &= \left|\int_a^b [f(x) - u_k(x)]dx\right| \\ &\leq \int_a^b |f(x) - u_k(x)|dx < \int_a^b \frac{\epsilon}{b - a} dx \leq \epsilon . \end{aligned}$$

Having demonstrated the importance of uniform convergence we need a usable criterion for establishing uniform convergence. This is provided, in the setting of series of functions, by the Weierstrass M-Test.

THEOREM 14-6c. (Weierstrass M-Test) Consider $u : n \to u_n$, where E is the domain of u_n for $n = 1 , 2 , \ldots$, and let $M : n \to M_n$ be a sequence of positive constants for which $|u_n(x)| < M_n$ for $x \in E$. If $\sum_{n=1}^{\infty} M_n$ converges, then $\sum_{n=1}^{\infty} u_n$ converges uniformly on E .

Proof. By the First Comparison Test $\sum_{n=1}^{\infty} u_n$ converges pointwise and absolutely, say to f . Since $\sum_{n=1}^{\infty} M_n$ converges, given $\epsilon > 0$, there exists an $\omega = \ (\epsilon)$ such that, for $k > \omega$

$$\left| \sum_{n=k+1}^{\infty} M_n \right| < \epsilon .$$

Thus if $k > \omega$, then

$$\begin{aligned} \left|f(x) - \sum_{n=1}^{k} u_n(x)\right| = \left|\sum_{n=k+1}^{\infty} u_n(x)\right| &\le \sum_{n=k+1}^{\infty} |u_n(x)| \\ &< \sum_{n=k+1}^{\infty} M_n < \epsilon \end{aligned}$$

for all $x \in E$.

Example 14-6c. The series $\sum_{n=0}^{\infty} x^n$ converges uniformly on $|x| < r$ where $r < 1$ since $|x^n| < r^n$ and $\sum_{n=1}^{\infty} r^n$ converges.

Example 14-6d. $\sum_{n=1}^{\infty} \frac{1}{n2^n} = \log 2$. By Theorem 14-6b

$$\log 2 = \int_0^{1/2} \frac{dx}{1 - x} = \sum_{n=0}^{\infty} \int_0^{1/2} x^n dx = \sum_{n=0}^{\infty} \frac{1}{(n + 1)2^{n+1}} = \sum_{n=1}^{\infty} \frac{1}{n2^n} .$$

THEOREM 14-6d. If the series of continuously differentiable functions $\sum_{n=1}^{\infty} u_n$ converges uniformly to f on the interval $I = (a,b)$ and if the series of derivatives $\sum_{n=1}^{\infty} u_n'$ converges uniformly to g on I, then f is differentiable and $f' = g$.

Proof. Since $\sum_{n=1}^{\infty} u_n' = g$ by Theorem 14-6b

$$\int_a^x g(t)dt = \sum_{n=1}^{\infty} \int_a^x u_n'(t)dt = \sum_{n=1}^{\infty} (u_n(x) - u_n(a)) = f(x) - f(a) .$$

Since g is continuous, $f'(x) = g(x)$.

Exercises 14-6

1. Show that each of the following series converges uniformly on the sets specified.

   (a) $\sum_{n=1}^{\infty} \frac{\sin nx}{n^2}$, $-\infty < x < \infty$;

   (b) $\sum_{n=1}^{\infty} \frac{\sin x^n}{n}$, $|x| < \frac{1}{2}$;

   (c) $\sum_{n=1}^{\infty} (\frac{x}{x+1})^n$, $0 < x < 2$.

2. Show that the Weierstrass M-Test is not a necessary condition for uniform convergence.

3. Show that $\sum_{n=0}^{\infty} x^n$ does not converge uniformly on $|x| < 1$.

4. In Example 14-6a and Example 14-6b show for each fixed $\epsilon < 1$, $\Omega(x,\epsilon)$ cannot be bounded.

5. A sequence of functions $u : n \to u_n$ is said to converge in the mean to f on $[a,b]$ if

$$\lim_{n\to\infty} \int_a^b [f(x) - u_n(x)]^2 dx = 0 .$$

   (a) Prove: if u converges uniformly to f on $[a,b]$ then u converges in the mean to f .

   (b) Show by an example that u can converge in the mean to f but not pointwise.

6. Show that if the series

(i) $\frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos nx + b_n \sin nx]$ converges uniformly to $f(x)$ on $[-\pi,\pi]$, then

(ii) $a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx$, $n = 0 , 1 , \ldots$

(iii) $b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx$, $n = 1 , 2 , \ldots$.

(The series (i) with coefficients defined by Equations (ii) and (iii) is called the Fourier **Series** of f .)

14-7. Power Series.

At the end of Section 13-3 we sought to determine when a function, f , could be the sum of its Taylor series $\sum_{n=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x - a)^k$. Section 8-5, Formula (8) showed that $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. These series are power series, series of the form $\sum_{n=0}^{\infty} a_n(x - a)^n$. To study power series it is sufficient to study power series of the special form

$$\sum_{n=0}^{\infty} a_n x^n . \tag{1}$$

Our first theorem implies the existence of an interval I symmetric about the origin such that (1) converges absolutely for all interior points of I and diverges for all exterior points. Thus, if I is bounded, $I = \{x : |x| < R\}$, (1) converges for $|x| < R$ and diverges for $|x| > R$. R is called the radius of convergence of (1). If I is unbounded, then I is the set R of all real numbers and we write $R = \infty$. That R can take on all values between 0 and infinity is shown by applying the root test to the following three examples.

1. $\sum_{n=0}^{\infty} (nx)^n \; (R = 0)$,
2. $\sum_{n=0}^{\infty} (ax)^n \; (R = \frac{1}{|a|})$ and
3. $\sum_{n=1}^{\infty} (\frac{x}{n})^n \; (R = \infty)$.

For $|x| = R$, the convergence of (1) is left in doubt as the following example shows.

Example 14-7a. $\sum_{n=1}^{\infty} \frac{x^n}{n}$ converges if $-1 \le x < 1$ and diverges otherwise.

THEOREM 14-7a. If the power series $\sum_{n=0}^{\infty} a_n x^r$ converges for $x = r$, then the power series $\sum_{n=0}^{\infty} a_n x^r$ converges uniformly and absolutely for $|x| \le r_1 < |r|$.

Proof. Since the series $\sum_{n=0}^{\infty} a_n r^n$ converges, $\lim_{n \to \infty} a_n r^n = 0$. Hence there is an ω such that $|a_n r^n| < 1$, for $n > \omega$. But, for $|x| \le r_1$, $|a_n x^n| = |a_n||x^n| \le |a_n||r_1^n| \le |a_n r^n||\frac{r_1}{r}|^n < |\frac{r_1}{r}|^n$ for $n > \omega$. Since $\frac{r_1}{r} < 1$, $\sum_{n=0}^{\infty} (\frac{r_1}{r})^n$ converges and $\sum_{n=0}^{\infty} a_n x^n$ converges uniformly for $|x| \le r_1$ by the Weierstrass M-Test.

From this result it is not hard to establish the existence of the radius of convergence mentioned above (Exercises 14-7, No. 2).

We have the following corollary to Theorem 14-7a.

Corollary. Let R be the radius of convergence of the power series $\sum_{n=0}^{\infty} a_n x^n$. If $R_1 < R$, the power series $\sum_{n=0}^{\infty} a_n x^n$ converges uniformly for $|x| < R_1$.

Thus by Theorem 14-6a if $f(x) = \sum_{n=0}^{\infty} a_n x^n$, the function f is continuous in $\{x : |x| < R\}$. Moreover, since the convergence of $\sum_{n=0}^{\infty} a_n r^n$, $r \ne 0$, implies

$$|n\ a_n x^{n-1}| \le |\tfrac{n}{r}|\,|a_n r^n| \left|\frac{r_1}{r}\right|^{n-1} \le |\tfrac{n}{r}| \left|\frac{r_1}{r}\right|^{n-1}$$

for $n > \omega$ and $|x| < r_1$, and since $\sum_{n=1}^{\infty} |\frac{n}{r}| \left|\frac{r_1}{r}\right|^{n-1}$ converges by the n-th root test, the power series $\sum_{n=1}^{\infty} n\ a_n x^{n-1}$ converges uniformly for $|x| < r_1 < r$. Hence by Theorem 14-6d $f'(x) = \sum_{n=1}^{\infty} n\ a_n x^{n-1}$. Thus, by induction f is infinitely differentiable and

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1) \ldots (n-k+1) a_n x^{n-k} . \tag{2}$$

Setting $x = 0$ in (2) we find $f^k(0) = k!\ a_k$ or

$$a_k = \frac{f^{(k)}(0)}{k!} . \tag{3}$$

Thus $\sum_{n=0}^{\infty} a_n x^n$ must be the Taylor Series of its sum f.

Exercises 14-7

1. If the series $\sum_{n=0}^{\infty} a_n x^n$ and $\sum_{n=0}^{\infty} b_n x^n$ converge on $|x| < R$ show that

$$\text{(i)} \quad \sum_{n=0}^{\infty} a_n x^n + \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} (a_n + b_n) x^n$$

and that

$$\text{(ii)} \quad \left(\sum_{n=0}^{\infty} a_n x^n\right)\left(\sum_{n=0}^{\infty} b_n x^n\right) = \sum_{n=0}^{\infty} c_n x^n$$

on any interval $|x| < p$ where $p < R$, where $c_n = \sum_{k=0}^{n} a_k b_{n-k}$.

2. From Theorem 14-7a prove the claim of the text that a power series $\sum_{n=0}^{\infty} a_n x^n$ either

(a) converges for all x , or

(b) there exists a number R such that the series converges for $|x| < R$ and diverges for $|x| > R$.

3. Prove that if $\sum_{n=0}^{\infty} a_n x^n$ has radius of convergence R_1 and $\sum_{n=0}^{\infty} b_n x^n$ has radius of convergence $R_2 < R_1$, then $\sum_{n=0}^{\infty} (a_n + b_n) x^n$ has radius of convergence R_2 .

4. Show that the radius of convergence of the power series $\sum_{n=0}^{\infty} a_n x^n$ is

$$R = \begin{cases} 0 & \text{if} \quad \overline{\lim}\ \sqrt[n]{|a_n|} = \infty \\ \infty & \text{if} \quad \overline{\lim}\ \sqrt[n]{|a_n|} = 0 \,, \end{cases}$$

in all other cases $R = \overline{\lim} \dfrac{1}{\sqrt[n]{|a_n|}}$.

5. Find the radius of convergence, R , for each of the following power series

(a) $\sum_{n=0}^{\infty} n(n+1)x^n$

(b) $\sum_{n=1}^{\infty} \frac{2^n x^n}{n}$

(c) $\sum_{n=0}^{\infty} \frac{n^k x^n}{n!}$

(d) $\sum_{n=0}^{\infty} \frac{n!\, x^n}{n^n}$

Miscellaneous Exercises

1. Extend the Second Comparison Test by proving that if $\lim_{n\sim\infty} \frac{b_n}{a_n} = 0$, where $a_n > 0$, for all n, and the $\sum_{n=1}^{\infty} a_n$ converges then $\sum_{n=1}^{\infty} b_n$ converges absolutely.

2. Let $\sum_{n=1}^{\infty} b_n$ be a convergent series of positive terms.

   Prove if $\left|\frac{a_{n+1}}{a_n}\right| < \frac{b_{n+1}}{b_n}$, for all n, then $\sum_{n=1}^{\infty} a_n$ converges absolutely.

3. Let $\sum_{n=1}^{\infty} a_n$ be a series of nonnegative terms.

   Prove if $n\left(\frac{a_n}{a_{n+1}} - 1\right) > 1 + \epsilon > 1$ for $n > \omega$, then $\sum_{n=1}^{\infty} a_n$ converges absolutely, but if $n\left(\frac{a_n}{a_{n+1}} - 1\right) < 1 - \epsilon < 1$ then $\sum_{n=1}^{\infty} a_n$ diverges.

   Hint: Use the preceding exercise to compare the given series with a p-series, where $p = 1 \pm \epsilon$.

4. Show that each of the conditions in Leibniz's Test (Theorem 14-4b) is necessary for convergence.

5. Prove that $\sum_{i=1}^{\infty} a_i$ converges if

   (a) $\operatorname{sgn} a_k = -\operatorname{sgn} a_{k+1}$,

   (b) $k \to |a_k|$ is nonincreasing,

   (c) $\lim_{k\sim\infty} a_k = 0$.

6. Show that if $\lim_{n\sim\infty} \left|\frac{a_{n+1} - a_n}{a_n - a_{n-1}}\right| = r$, then the sequence $n \to a_n$ converges if $r < 1$ and diverges if $r > 1$.

7. Translate the Weierstrass M-Test as a criterion for the uniform convergence of a sequence of function $u : n \to u_n$.

8. If, for all x in E , $|u_\nu(x)| \leq M$ and $\left|\frac{u_{n+1}(x)}{u_n(x)}\right| < r < 1$ for all $n \geq \nu$, then $\sum_{n=1}^{\infty} u_n$ converges uniformly in E .

9. A telescoping series is a series of the form $\sum_{n=1}^{\infty} (a_n - a_{n+1})$. Give necessary and sufficient conditions for the convergence of a telescoping series.

10. Prove the Cauchy Criterion for uniform convergence: a necessary and sufficient condition for the uniform convergence of the sequence of functions $u : n \to u_n$ with common domain E is that to every $\epsilon > 0$ there exists an $\omega = \Omega(\epsilon)$ such that if $n , m > \omega$, then $|u_n(x) - u_m(x)| < \epsilon$ for all x in E .

11. Show if the series of functions $\sum_{n=1}^{\infty} v_n$ converges uniformly on E and the series of functions $\sum_{n=1}^{\infty} u_n$, with common domain E , has the property that $|u_n(x)| \leq v_n(x)$ for all $x \in E$, then $\sum_{n=1}^{\infty} u_n$ converges uniformly in E .

12. (a) Consider a series of functions $\sum_{n=1}^{\infty} u_n$ uniformly convergent to U on E . Let f be a function defined and bounded on E , $|f(x)| \leq M$. Prove that $\sum_{n=1}^{\infty} f \cdot u_n$ converges uniformly to

$$f \cdot \sum_{n=1}^{\infty} u_n = U .$$

(b) Show, by an example, that the boundedness of f is a necessary condition in Part (a).

13. Find the Taylor expansion of $f : x \to (1 + x)^{\alpha}$, α not a positive integer, (the binomial series for exponent α) , and find its radius of convergence.

14. Show that the radius of convergence, R , of the Taylor series of arc sin x ,

$$\sum_{k=0}^{\infty} \frac{(2k)!t^{2k+1}}{(2k + 1)(k!)^2 \, 2^{2k}} ,$$

(see Example 13-3b) is 1 .

15. Show that if the continuous function $(x,y) \to \Phi(x,y)$ defined in the rectangle

$$\{(x,y) : |x - x_0| \leq a , |y - y_0| \leq c\}$$

satisfies $ab < 1$, $a\Lambda \leq c$, where

(1) $\max\{|\Phi(x,y)| : |x - x_0| \leq a , |y - y_0| \leq c\} = \Lambda$

and

(2) $\max\{|D_y \Phi(x,y)| : |x - x_0| \leq a , |y - y_0| \leq c\} = b$,

then the sequence of functions $u : k \to u_k$ defined by $u_0 : x \to y_0$,

$$u_{k+1}(x) = \int_{x_0}^{x} \Phi(x,u_k(x))dx$$

converges to a function U which satisfies the differential equation

$$\frac{dy}{dx} = \Phi(x,y) \quad \text{for} \quad |x - x_0| < a ,$$

and the initial condition $y = y_0$ at $x = x_0$.

M16. Find the value of $a^{a^{a^{\cdot^{\cdot^{\cdot}}}}}$. More precisely, find the limit of the sequence $n \to x_n$ defined by $x_0 = a$, $x_{n+1} = a^{x_n}$ and determine the values of a for which the sequence converges.

Chapter 15

GEOMETRICAL OPTICS AND WAVES

15-1. Introduction.

This is the third chapter on applications of the calculus to the sciences. In the first of these, Chapter 9, we were concerned with processes of growth, decay, and competition, and showed that one differential equation may govern natural phenomena in many different and, at first, seemingly unrelated contexts. Our main purpose was to stress how one mathematical model may serve to link phenomena and processes which occur in all the sciences. In the second of these chapters, Chapter 12, we pursued the one science of Mechanics to some depth, but again followed a narrow mathematical thread; the solution of certain differential equations (primarily those of oscillatory phenomena) was our basic guide. There we saw how a few basic mathematical ideas can be exploited for the analysis of a varied assortment of problems arising in a single science. We selected a narrow thread of mathematical methods to weave across the sciences.

The intent of the present chapter is quite different. Now we use a science as the guiding thread: we select a narrow sequence of physical concepts leading from geometrical optics through wave physics. The development of the science will use most of the important methods of the calculus. Our purpose is to show how the broad sweep of the calculus (in contrast to the limited selection of methods in Chapters 9 and 12) finds application, and we shall see that the calculus furthers the development of the science at every stage.

Since our main concern here is mathematics and not physics, we shall not try to lay grounds for belief in physical laws. We simply accept "laws of nature" that were isolated only after long years of observation, speculation, and verification, and reveal the consequences implicit in these laws by methods of calculus. As we progress along the scientific thread, we trace part of the broad line of historical development from the early laws of geometrical optics with their limited domain of applicability to the modern very general laws of wave theory.

Chapter 9 is intended to show how one topic of mathematics is universally applicable to all the sciences. This chapter is intended to show how one topic of science uses various mathematical methods. The two chapters together are intended to indicate that the interactions of mathematics and science are profound indeed. Mathematics and science are the very warp and woof of the conception of the universe that we have created. Where mathematics and science are most highly interwoven, there, both subjects have tended to reach their greatest development.

15-2. Geometrical Optics.

We introduce physical laws in much the order in which they were set forth. We start with very restrictive laws, weaken the restrictions, and thereby enlarge the domain to which the laws apply. The procedure we follow is "quasi-axiomatic" and is intended to suggest both the physicist's heuristic search for fundamental principles on the basis of limited observations as well as his tests of such principles by further observations. The laws or "axioms" we list contain undefined terms as do the axioms of a mathematical system, but they also necessarily contain implicit restrictions on both the observational and computational procedures that are associated with them. In mathematics, too, the axioms are usually stated in an implicit context, but the situation is more obvious in science.

We define neither light, nor its various subjective attributes: we start with "let there be light", and introduce mathematical constructs that represent measurable properties of light. You are aware that you are reading with the aid of light, and although this page is at present practically fully illuminated, you can change the situation by closing the book (something we writers try always to bear in mind). There is no obvious structure to the flood of illumination upon the page, but you have seen dancing spots of light traced by shafts of sunlight on a tree-shaded path, and beams of light entering darkened rooms through narrow cracks of slightly open doors; or, you may have become aware of straight line characteristics revealed in floods of sunlight by the shadows that they cast. You have handled light sources such as electric lights, flashlights, candles, and have seen the stars as distant sources of light. You have seen the image of your face in a mirror, and the fractured appearance at the surface of water of partially immersed objects.

The most primitive constructs for such situations are geometrical, and were introduced by Euclid. Euclid represented light as something "propagating" (traveling) along "rays" (straight lines) and "reflected" (thrown back in a special way) when it encountered a mirror. A geometry of rays and the hypothesis that light travels at different speeds in different transparent materials serve also to account for the "refraction" (breaking or bending) of a ray passing from, say, air to water.

Why introduce the idea of traveling? The candle that is consumed as it gives forth light, and the monthly electric bill, make clear that something is being used up to provide the light. We transform energy from some other form to the form associated with light; the energy flows from one place to another and the rays, we consider, are guides for the flow of energy.

In the next two sections we apply the calculus only to the geometry of rays -- mostly straight rays, but also some curved ones; mostly familiar effects, but not necessarily familiar interpretations: we introduce a signed ray (a "shadow forming ray") to account for shadows, and some of the rays may split into many ("diffraction") to describe some aspects of what occurs when light strikes a sharp edge. Sections 15-2 and 15-3 deal with geometry, and geometry will be sufficient until we associate the magnitude of energy flow with a ray (as we do in Section 15-4). However, to appreciate the physical content, it should be kept in mind that these rays, in some sense, are the directions for energy flow.

(i) Euclid's Principles.

Early observations of light sources (sun, stars, lamps) and of the reflections of such sources and objects in smooth surfaces (water, polished metal) suggested two "laws" of nature; Euclid's principles of propagation [E1] and of reflection [E2]:

[E1]: light travels along straight lines (rays);

[E2]: when a ray is incident on a smooth plane surface, the incident ray, the reflected ray, and the normal to the surface all lie in the same plane, and the two rays make equal angles on the opposite sides of the normal.

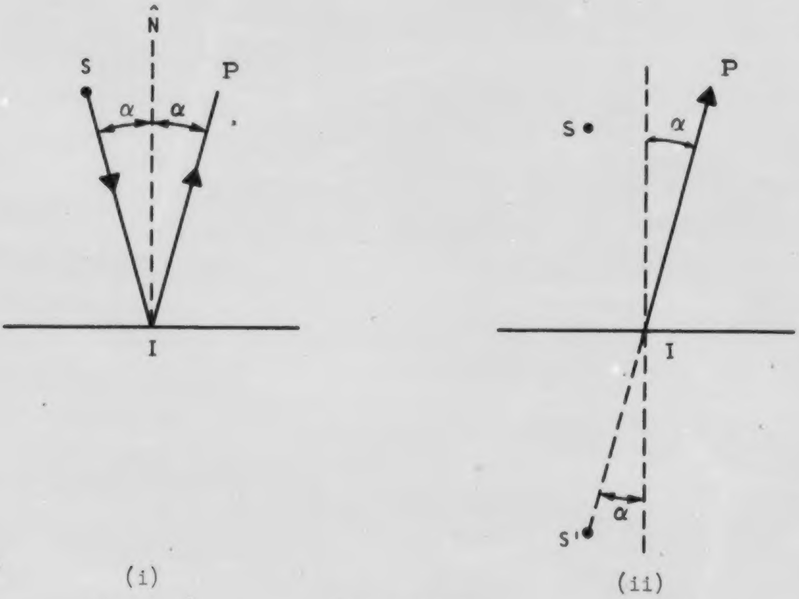
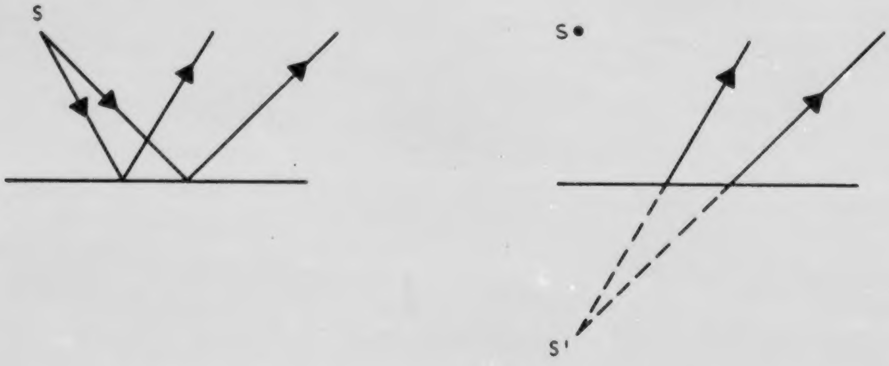


Figure 15-2a

Figure 15-2a(i) illustrates [E2] ; it shows the plane of incidence in which a ray from a source (S) reaches the observation point (P) via reflection at the intersection point (I) on the mirror; the rays are at angles α with the surface normal (N) . In Figure 15-2a(ii), we see that the reflected ray (I to P) is the extension of the mirror image (S' to I) of the incident ray (S to I).



(i) (ii)

Figure 15-2b

If a set of rays diverging from a source S is reflected from a plane mirror as in Figure 15-2b(i), the corresponding set of reflected rays appear to originate from the image source S' as in Figure 15-2b(ii). Thus as far as the reflected set of rays (reflected ray system) is concerned, we may replace the mirror in Figure 15-2b(ii) by the source S' and reduce a reflection problem specified by [E2] to a propagation problem specified by [E1] . (This image method was essentially introduced by Heron or Hero several hundred years after Euclid.)

We regard [E1] as defining geometrical propagation in a uniform medium, and [E2] as defining geometrical reflection from smooth planes. These cover the situation of Figures 15-2a and 15-2b as well as more complicated reflection problems.

Two parallel rays incident on a planar reflector as in Figure 15-2c(i) are reflected as parallel rays. If we regard reflector as consisting of two hinged planes, and swing one away from the source as in Figure 15-2c(ii), then reflected rays are said to diverge; if instead, we swing the plane toward the source, then the reflected rays converge as in Figure 15-2c(iii). In the latter case, the reflected rays intersect, while in Figure 15-2c(ii), their extensions "behind" the mirrors intersect. When the rays converge, their intersection is called a real intersection, and when they diverge, the intersection of their extensions is called virtual. In either case, the reflected rays appear to originate at the intersection.

If we have rays incident on a complex reflector consisting of many planar portions, then we can determine the reflected rays by applying [E2] . Equivalently, once we have determined the intersections (real or virtual) of the reflected rays we have reduced the problem to an application of [E1] with the intersection point taken as the image source. We also need to determine the intersections of rays reflected from curved mirrors. But before we can consider curved mirrors, we must introduce a more general law of nature than [E] .

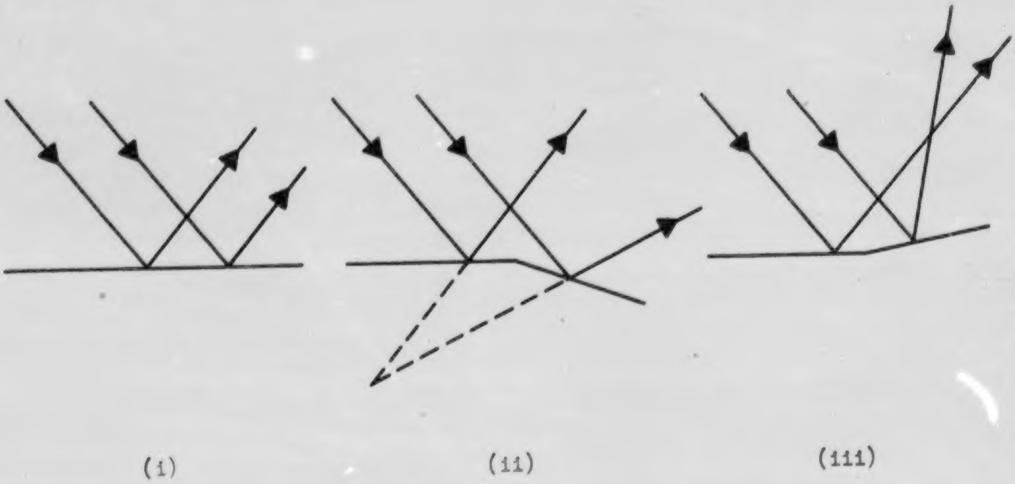

(i) (ii) (iii)

Figure 15-2c

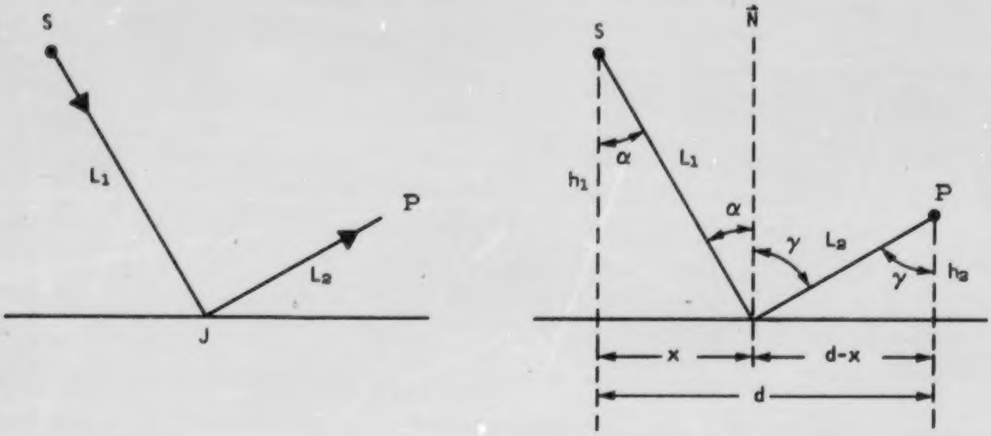
(ii) Hero's Principle.

Euclid's principles, which describe the essentials of many directly observable phenomena, are contained in the more concise and more general principle of Hero:

[H] : a ray follows the shortest path between points.

With Hero's principle we may study reflection by curved mirrors.

Before applying [H] to more general situations than those covered by [E] , we use the calculus to derive [E] from [H] . Since, by definition, a straight line is the shortest path between points, we see that [H] covers [E1] directly. Similarly in considering [E2] we need not consider curved paths. We seek the shortest path between S and P via a point J on the surface of the mirror as in Figure 15-2d(i). We may take J in the plane through S and P that is perpendicular to the mirror: any displacement of J perpendicular to this plane will clearly lengthen the component paths L_1 and L_2 (Exercises 15-2, No. 1). We introduce the lengths and angles of Figure 15-2d(ii) and minimize $L = L_1 + L_2$ (as required by [H]) and show that the minimum corresponds to $\gamma = \alpha$ (as required by [E2]).



(i) (ii)

Figure 15-2d

In order for

$$L = L_1 + L_2 = \sqrt{h_1^2 + x^2} + \sqrt{h_2^2 + (d - x)^2} \tag{1}$$

to be a minimum for fixed h_1 and h_2, and J on the surface of the mirror, we require

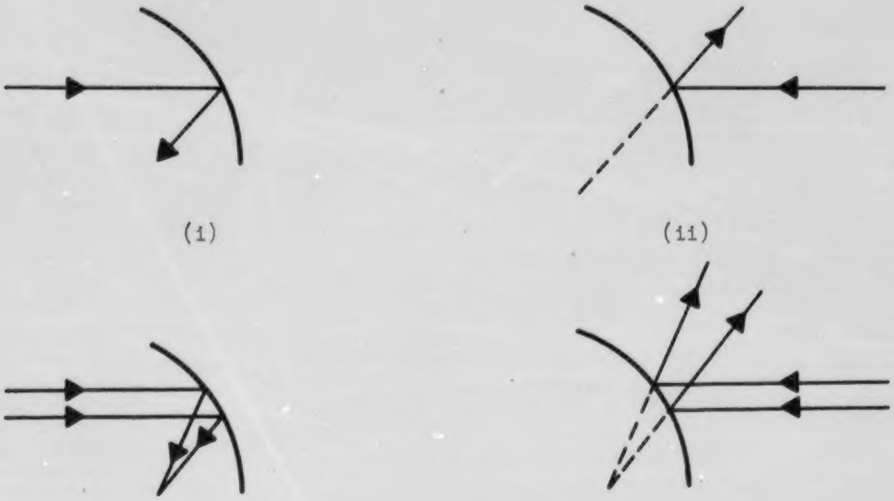
$$\frac{dL}{dx} = 0 . \tag{2}$$

Thus

$$\frac{dL}{dx} = \frac{x}{\sqrt{h_1^2 + x^2}} - \frac{(d - x)}{\sqrt{h_2^2 + (d - x)^2}} = \frac{x}{L_1} - \frac{d - x}{L_2} = \sin\alpha - \sin\gamma = 0 ,$$

and consequently $\sin\gamma = \sin\alpha$ as in [E2]. (See Exercises 15-2, No. 2.)

It is clear from the above that applying [H] to reflection from a point on a curved surface is equivalent to using [E2] for reflection from the tangent plane at the point. Practical applications prior and subsequent to [H] have been based on [E2] plus the "tangent plane approximation". Figure 15-2e(i) shows reflection of a ray from a point on the concave side of a reflector, and Figure 15-2e(ii) shows reflection at the same point of a ray



(i) (ii)

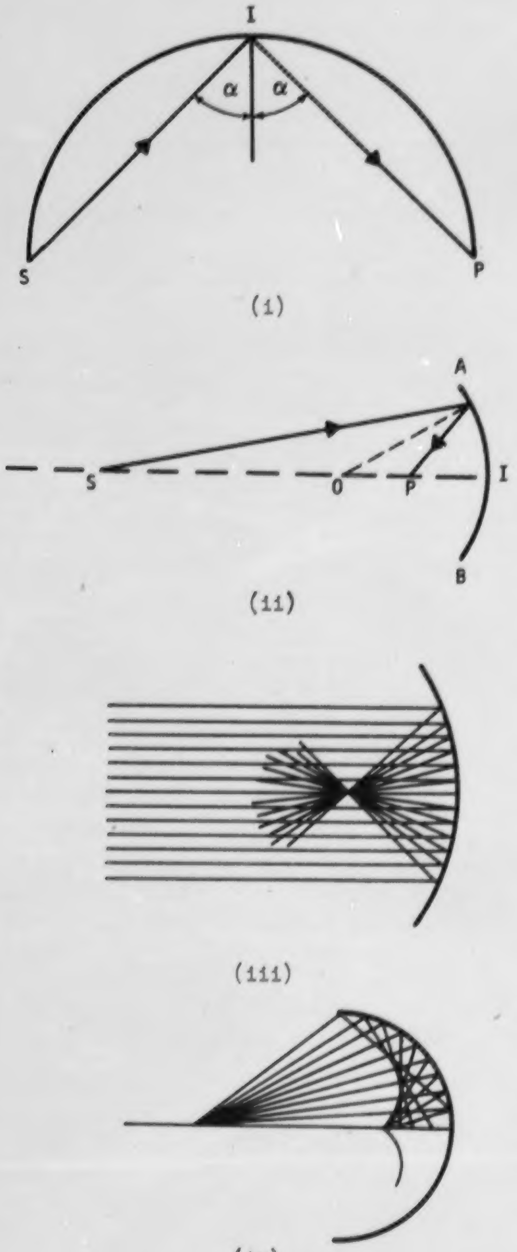
(iii) (iv)

Figure 15-2e

arriving from the convex side; note the relation between the directions. Similarly for two rays incident on a curved mirror, we can construct the reflected rays (or equivalently, their intersection) by geometry; note that both situations in Figures 15-2e(iii) and (iv) give rise to the same intersection point (real for (iii) and virtual for (iv)).

There are situations not covered by [H] which are covered by [E2] and the tangent plane approximation. For example, consider a source (S) on the circumference of a circle and an observation point (P) at opposite ends of a diameter. The geometrically reflected ray from S to P via a point I on the circumference as in Figure 15-2f(i), is the longest of all such paths (Exercises 15-2, No. 3). Equation (2) is fulfilled for the situation of Figure 15-2f(i), but for this case the corresponding second derivative is less than zero. There are also situations of interest covered by (2) for which the second derivatives vanish. For all such cases, independently of the second derivative, the first derivative of the path length is zero, and we say that the ray path is stationary for first order variations. To cover all such situations, we replace [H] by the more general principle

[H'] : a ray follows a stationary path.



(i)

(ii)

(iii)

(iv)

Figure 15-2f

We may distinguish two classes of curved reflectors and illustrate the essentials for the case of a concave cylindrical mirror. The simpler class corresponds to a "small aperture" mirror as in Figure 15-2f(ii); for this case the semi-aperture of the mirror $\frac{|AB|}{2}$ is very small compared to its radius of curvature a. Let I be the center of the arc $\widehat{AB}$ and S the position of the source on a line perpendicular to the mirror at I; then from the law of reflection, it follows that to a first approximation the rays reflected from all points of the mirror go through the point P for which

$$\frac{1}{|SI|} + \frac{1}{|PI|} = \frac{2}{a} .$$

The proof is left to Exercises 15-2, Number 4. In particular if $|S| \sim \infty$, then the situation corresponds to an incident set of parallel rays as in Figure 15-2f(iii), and we obtain simply

$$|PI| = \frac{a}{2} .$$

Thus for the small-aperture mirror all reflected rays intersect at $\frac{a}{2}$ (the focus).

The more general problem of reflection from a mirror with arbitrary sized aperture is illustrated in Figure 15-2f(iv) for a set of rays from a source on the axis of a semi-circular mirror (Figure 15-2f(ii) depicts only the situation in the vicinity of the axis). If we rotate this figure around its symmetry axis, we obtain a representation of reflection from a portion of a spherical mirror. If the distance of the source from the reflector becomes infinite (parallel set of rays incident), then the envelope of the reflected rays is an epicycloid (Exercises 11-M, No. 5); the cusp of this curve is at $\frac{a}{2}$. The envelope of these reflected rays is called a caustic; we may deal with a caustic surface, a caustic line, or a caustic point; the last is also called a focus. Next we consider the caustic obtained by the reflection of a set of parallel rays incident upon a cylindrical mirror. By restricting the rays to a plane perpendicular to the generators of the cylinder we obtain a two-dimensional problem. In this case we expect to obtain a caustic line. Since the reflected rays are tangent to the caustic, once we know the caustic we specify the system of reflected rays by means of [E1].

(iii) Caustics.

We consider a set of rays parallel to the x-axis incident upon a two-dimensional mirror. We consider a fixed observation point $P = (x,y)$ on the same side of the reflector as the source, and apply [H] to determine the condition that the incident ray that strikes the mirror at $I = (\xi, \eta)$ is reflected through P (see Figure 15-2g). We could specify the point I and

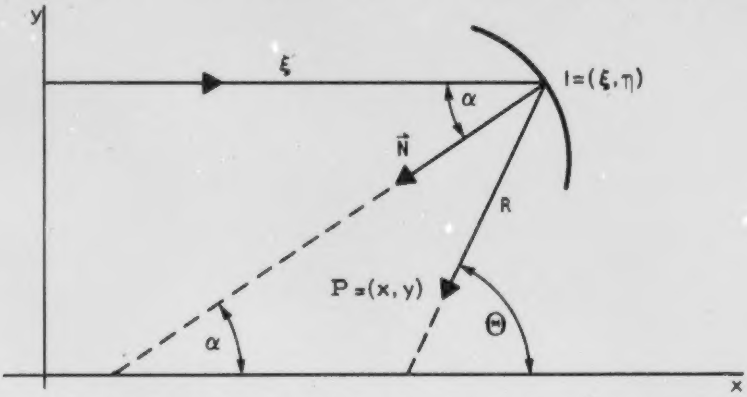


Figure 15-2g

the incident ray that strikes it in terms of the parameter of arc length along the curve, but it is more convenient to use the angle that the incident ray makes with the normal at I as the parameter, namely the angle α such that $\tan\alpha$ is the slope of $\vec{N}$. The length of the incident ray measured from the y-axis is ξ; the length of the reflected ray from I to P is $R = \sqrt{(\xi - x)^2 + (\eta - y)^2}$, and its inclination to the x-axis is θ.

The total length from the y-axis via I to P equals

$$L = \xi + R = \xi + \sqrt{(\xi - x)^2 + (\eta - y)^2}\,. \tag{3}$$

Differentiating, we have

$$L' = \xi' + \frac{\xi'(\xi - x) + \eta'(\eta - y)}{R} = \xi'[1 + \cos\theta] + \eta'\sin\theta \tag{4}$$

where the prime indicates differentiation with respect to α. Using [H], essentially as for (2), we equate L' to zero to obtain

$$\frac{-\xi'}{\eta'} = \frac{\sin\theta}{1 + \cos\theta} = \frac{2\sin\frac{\theta}{2}\cos\frac{\theta}{2}}{2\cos^2\frac{\theta}{2}} = \tan\frac{\theta}{2}\,. \tag{5}$$

Thus we have

$$\tan\frac{\theta}{2} = \frac{-\xi'}{\eta'} = \frac{-d\xi}{d\eta}\,. \tag{6}$$

Since $\frac{d\eta}{d\xi}$ is the slope of the tangent of the reflector at the point of incidence, $\frac{-d\xi}{d\eta}$ is the slope of the normal $\vec{N}$ and equals $\tan\alpha$. Consequently

$$\frac{-\xi'}{\eta'} = \tan\theta = \tan\alpha \ , \tag{7}$$

from which

$$\theta = 2\alpha \ , \tag{8}$$

as could have been obtained directly from [E] upon inspection of Figure 15-2g.

The equation of the reflected ray arising from the ray incident at an angle α with $\vec{N}$ is

$$\eta - y = (\xi - x)\tan\theta = (\xi - x)\tan 2\alpha \ , \tag{9}$$

which we rewrite as

$$g(\alpha,P) = (\xi - x)\tan 2\alpha - (\eta - y) = 0 \ . \tag{10}$$

This specifies the set of reflected rays corresponding to a set of incident parallel rays. The parameter α describes not only the curve of the reflector $(\xi(\alpha), \eta(\alpha))$, it also picks out the ray incident at ξ, η and the corresponding reflected ray (10). The point of intersection of two neighboring rays $g(\alpha,P) = 0$ and $g(\alpha + \Delta\alpha, P) = 0$, corresponding to α and $\alpha + \Delta\alpha$, is determined by $g(\alpha,P) = 0$ and $\dfrac{g(\alpha + \Delta\alpha, P) - g(\alpha,P)}{\Delta\alpha} = 0$. In the limit $\Delta\alpha \to 0$, the point P of intersection of the neighboring rays falls on the envelope and is specified by the simultaneous equations

$$g(\alpha,P) = 0 \quad , \qquad D_\alpha\, g(\alpha,P) = 0 \ . \tag{11}$$

Differentiating (10), we obtain

$$D_\alpha g(\alpha,P) = \xi' \tan 2\alpha + \frac{2(\xi - x)}{\cos^2 2\alpha} - \eta' \ , \tag{12}$$

and with (7) we eliminate $\eta' = \dfrac{-\xi'}{\tan\alpha}$:

$$D_\alpha g(\alpha,P) = \xi'\left(\tan 2\alpha + \frac{1}{\tan\alpha}\right) + \frac{2(\xi - x)}{\cos^2 2\alpha} \ . \tag{13}$$

Since $\tan 2\alpha \tan\alpha + 1 = \dfrac{1}{\cos 2\alpha}$, we reduce (13) to

$$D_\alpha g(\alpha,P) = \frac{1}{\cos^2 2\alpha}\left(\frac{\xi' \cos 2\alpha}{\tan\alpha} + 2(\xi - x)\right) \ . \tag{14}$$

Thus, since (11) requires $D_\alpha g(\alpha,P) = 0$, the x-coordinate of the point on the envelope is

$$x = \xi + \frac{\xi' \cos 2\alpha}{2\tan\alpha} \ , \tag{15}$$

which we may rewrite in various equivalent forms, e.g.,

(16) $$x = \xi - \frac{\eta'}{2}\cos 2\alpha \ .$$

We obtain the corresponding y-coordinate by using (16) to eliminate $\xi - x$ from (10): thus $g(\alpha,P) = \frac{\eta'}{2}\cos 2\alpha \tan 2\alpha - (\eta - y) = 0$, and consequently

(17) $$y = \eta - \frac{\eta'}{2}\sin 2\alpha \ .$$

Equations (15) and (17) (from which we could eliminate α) specify the caustic curve, the envelope of the reflected rays. Given a specific reflector we can use its parametric representation to eliminate ξ and η , and thereby determine the caustic explicitly. We illustrate this for two simple reflectors, the parabola and semi-circle.

Example 15-2a. Parabola. We consider a set of rays incident upon the parabola

(18) $$\eta^2 = -4p\xi$$

as in Figure 15-2h. The parametric equations of the parabola in terms of α ,

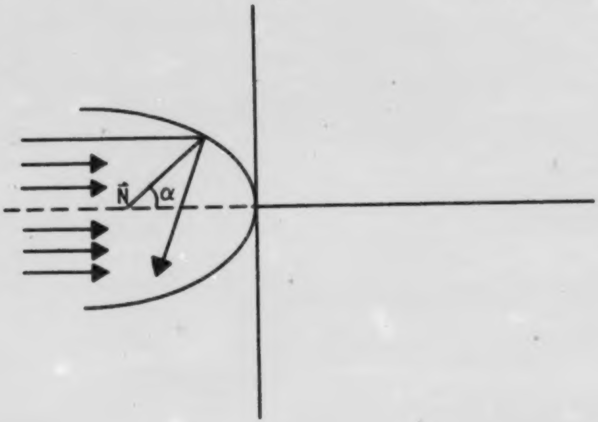


Figure 15-2h

where $\tan\alpha$ is the slope of the normal, are

(19) $$\eta = 2p\tan\alpha \quad , \quad \xi = -p\tan^2\alpha \ ,$$

and consequently

(20) $$\eta' = \frac{2p}{\cos^2\alpha} \quad , \quad \xi' = -p\,\frac{2\tan\alpha}{\cos^2\alpha} \ .$$

Entering the expressions for η and η' in (17), we find

$$(21) \qquad y = 2p \tan\alpha - \frac{p}{\cos^2\alpha}\sin 2\alpha = 2p\left(\tan\alpha - \frac{\sin\alpha}{\cos\alpha}\right) = 0 .$$

Similarly, employing the expressions for ξ and ξ' in (15), we have

$$(22) \quad x = -p \tan^2\alpha + \left(\frac{-p\ 2\tan\alpha}{\cos^2\alpha}\right)\left(\frac{\cos 2\alpha}{2\tan\alpha}\right) = \frac{-p}{\cos^2\alpha}(\sin^2\alpha + \cos 2\alpha) = -p .$$

Thus the envelope of the reflected rays is the point

$$(23) \qquad x = -p \quad , \quad y = 0 ,$$

namely, the focus of the parabola (see Figure 15-2i). The focusing property of the parabola accounts for its many applications (telescope mirrors, microwave and sonic "dishes", etc.) for collecting practically parallel radiation (the rays from very distant sources) by reflecting the incident rays to an appropriate small detector placed at its focus. Similarly, parabolic reflectors are used for the inverse problem of converting the radiation from a source at the focus into a parallel beam of rays.

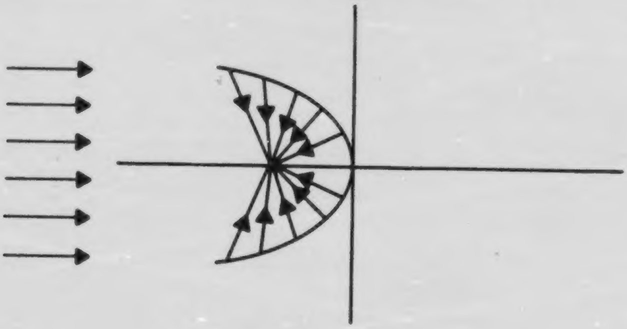


Figure 15-2i

The preceding example may seem artificial in that (23) could have been obtained by much simpler procedures than the one we followed. In the next example the same procedure yields a far less obvious result.

Example 15-2b. Semicircle. The parametric equations of a circle of radius a as depicted in Figure 15-2j are

$$(24) \qquad \eta = a\sin\alpha \quad , \quad \xi = a\cos\alpha ,$$

and consequently

$$(25) \qquad \eta' = a\cos\alpha \quad , \quad \xi' = -a\sin\alpha .$$

From (17) we then obtain for the caustic

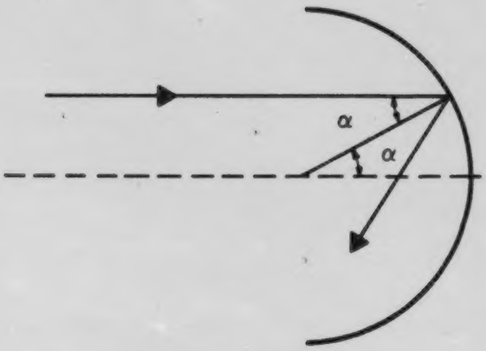


Figure 15-2j

(26) $$y = a \sin \alpha - a \cos \alpha \, \frac{\sin 2\alpha}{2} = a \sin^3 \alpha \ ,$$

and from (15),

(27) $$x = a \cos \alpha - \frac{a \sin \alpha \cos 2\alpha}{2 \tan \alpha} = \frac{a \cos \alpha}{2}(1 + 2 \cdot \sin^2 \alpha) \ .$$

Squaring in (26) and (27) and adding, we obtain

(28) $$\frac{4}{a^2}(x^2 + y^2) = 1 + 3 \sin^2 \alpha = 1 + 3\left(\frac{y}{a}\right)^{2/3} \ ,$$

the equation of the epicycloid traced out by a point on a circle of radius $\frac{a}{4}$ rolling on the outside of a fixed circle of radius $\frac{a}{2}$.

The cusp or focal point of the caustic is at $x = \frac{a}{2}$, $y = 0$; this corresponds to $D_\alpha^2 \, g(\alpha,P) = 0$, and occurs at $\alpha = 0$. The rays incident near the center of the mirror $(\alpha \approx 0)$ are known as paraxial rays of "small aperture" mirror theory; only these give rise to reflected rays that appear to originate at the cusp $\frac{a}{2}$. (See Exercises 15-2, No. 5.)

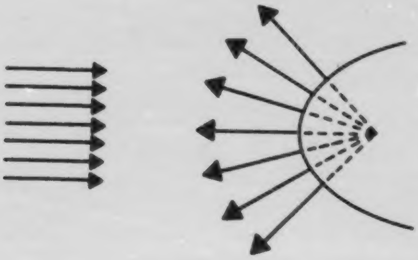
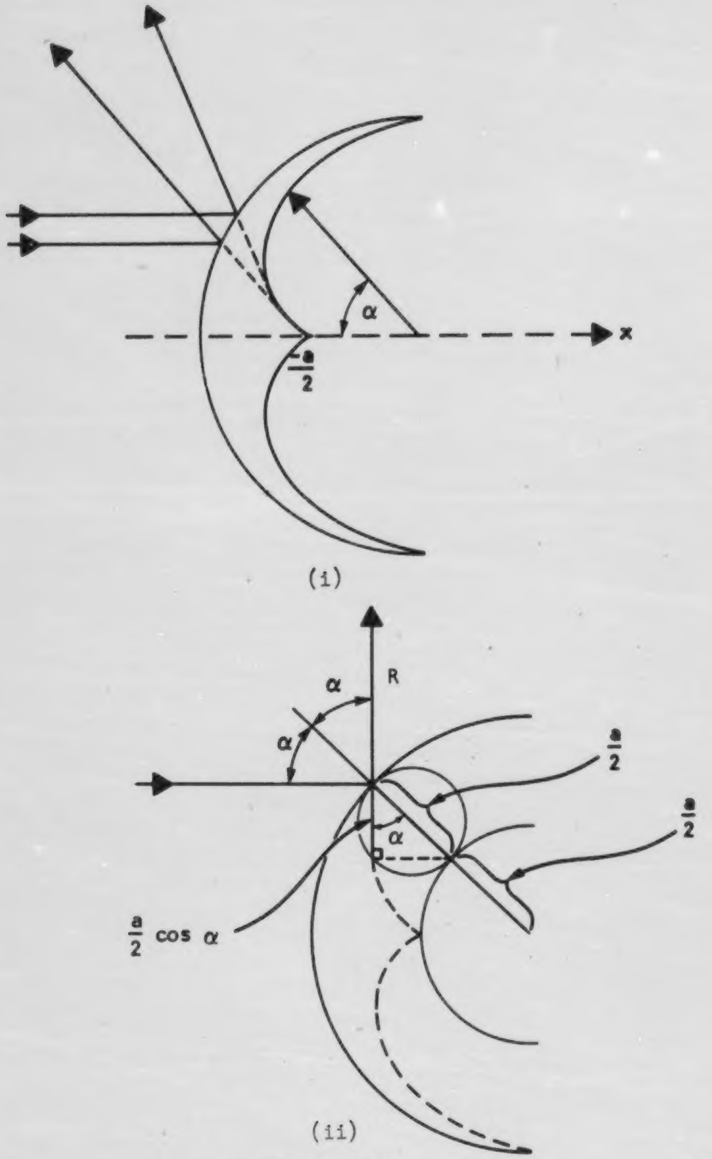


Figure 15-2k



(i)

(ii)

Figure 15-2ℓ

So far we have considered only reflections from concave mirrors; for such cases the reflected rays intersect and the caustics are real in the sense defined in Section 15-2(ii). Similarly, for incidence on a convex reflector the extensions of the reflected rays behind the reflector intersect on a virtual caustic. The identical caustic curve specifies reflection from either side of the mirror. Figure 15-2k shows the situation for incidence on a convex parabolic reflector, and Figure 15-2ℓ(i) shows the analogous situation for a semicircle. Figure 15-2ℓ(ii) shows the geometrical method of constructing the epicycloidal caustic of the semicircle. Since the caustic of Figure 15-2ℓ(ii) is the envelope of the set of extended reflected rays, it is tangent to all members of the family. From the figure we see that the distance along the ray extension from the mirror to the point of tangency with the caustic equals $\frac{a}{2}\cos\alpha$. Thus the neighboring reflected rays of length R appear to diverge from a source (their point of intersection) at a distance $R + \frac{a}{2}\cos\alpha$ along their extension.

Since the reflected rays are tangent to the caustic, we may treat the caustic as the evolute of a system of curves which are orthogonal to the rays. These curves, the involutes of the caustic, are called the eikonals or eikonal curves in ray theory; the radius of curvature at a point P on such a curve equals $R + \frac{a}{2}\cos\alpha$ where R is the distance along the ray from the mirror (Exercises 15-2, No. 7). The rays (the orthogonal trajectories of the eikonal curves) are tangent to the caustic and normal to the eikonals, and this provides a geometrical construction for the eikonals: they are traced by the points of a taut string as it unwinds from the caustic.

(iv) Shadows.

In the preceding discussion we took an observation point P lying on the same side of the reflector as the source (the "lit side" of the reflector). If we allow P to lie on the opposite side of the reflector, we obtain an additional solution of $L' = 0$ with L' as given in (4), i.e.,

$$L' = 0 \quad \text{if} \quad \theta = \pi \,, \tag{29}$$

where the geometry is shown in Figure 15-2m. Thus in addition to the geometrically reflected ray shown in Figure 15-2g, we see from $L' = 0$ (i.e., from [H']) that the incident ray also gives rise to another ray -- one traveling along the original direction of incidence. Were the reflector absent, we would interpret this ray as the incident ray itself (i.e., the situation of [E1]). However, we insist on the presence of the reflector and seek a physically significant interpretation of the rays corresponding to (29). When we interrupt a broad beam of light by a mirror, we notice essentially two effects:

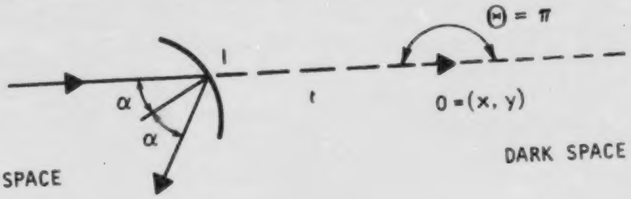


Figure 15-2m

because of the mirror, there is not only some light observed in a region of space outside of the original beam, but there is also some light missing from a region of space originally filled by the beam before we inserted the obstacle. Were we interested solely in the original beam, then we might simply say that some of the light has been "bent" from its original direction (reflected) and let it go at that. However in order to ultimately specify the full effect of the obstacle analytically, we assign it a more positive role. We say that the incident rays "excite" the obstable to produce not only the set of reflected rays but also a set of shadow forming rays parallel to the "missing" incident rays in the dark region of space. It is these shadow forming rays that we read into (29); these must cancel the incident rays on the "dark side" of the mirror to "create" the geometrical shadow. (This idea of shadow forming rays may be hard to reconcile with mental images of the reflection of rays based on a ball bouncing off a wall. However, were we interested in specifying the total effect of the wall in the ball-wall problem, we could also do so in terms of reflected trajectories and shadow forming trajectories.)

To make the full effect of the obstruction more explicit (and to set the stage for our subsequent discussion of scattering) we introduce a symbolic representation for the rays. We let E_i be a measure* at any point of the effect of an incident ray, E of the geometrically reflected ray, and E_s of the shadow forming ray. We represent the total effect E_t at any point corresponding to a ray E_i incident on a reflector, by

$$E_t = E_i + E : E = \begin{cases} E_g \text{ , in lit space} \\ E_s \text{ , in dark space .} \end{cases} \tag{30}$$

Thus in the lit space the total effect is $E_t = E_i + E_g$ as shown by the two rays on the left-hand side of Figure 15-2m. On the other hand in the dark space we have $E_t = E_i + E_s$ corresponding to the dashed ray on the right-hand side of Figure 15-2m; in order that E_t represent the physical situation of the geometrical shadow, i.e., in order that E_t vanish, we require

$$E_s = -E_i \text{ .} \tag{31}$$

We take (31) as a supplementary assumption to [H'] : the first solution $(\theta = 2\alpha)$ of $L' = 0$ accounts for geometrical reflection (and we subsequently determine a magnitude to be assigned to such rays); the second solution $(\theta = \pi)$ with (31) accounts for shadow formation.

---

*This measure will later be identified with the idea of "amplitude."

The symbol E in (30) represents the scattered part of the total effect $E_t = E_i + E$. This is the part of E_t that we may regard as originating at the obstacle to E_i, or as outgoing from the obstacle.

If we consider a system of parallel rays incident on a convex semi-circular cylinder (or equivalently on a full circular cylinder), then the corresponding scattered ray system (reflected plus shadow forming rays) is as sketched in Figure 15-2n.

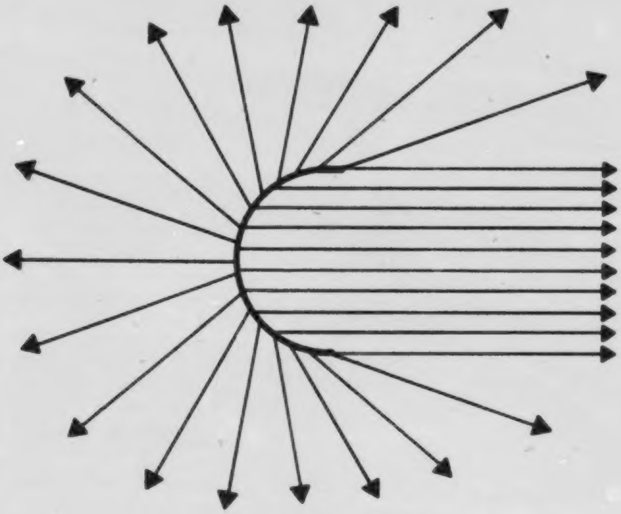


Figure 15-2n

The family of curves perpendicular to these rays is the corresponding infinite set of eikonals. Figure 15-2o plus its reflection in the x-axis shows several of these curves. These curves may be obtained geometrically from the caustics (the caustic for the shadow forming rays is the point at $x = -\infty$), or by constructing the normals of Figure 15-2n geometrically, or analytically. We give an analytical derivation in a following section. At larger and larger distances from the scatterer the eikonals of Figure 15-2o become more and more circular.

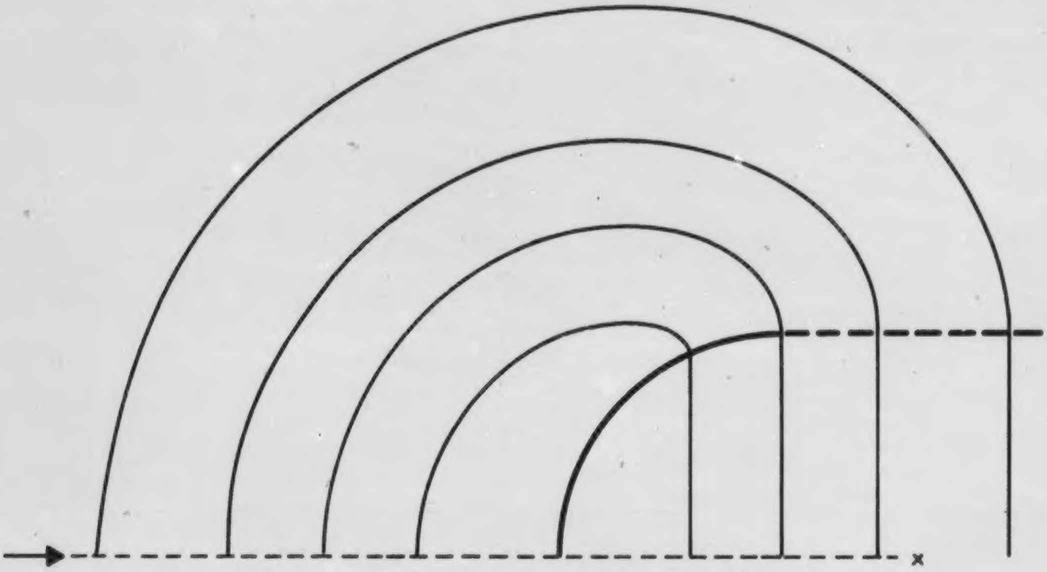
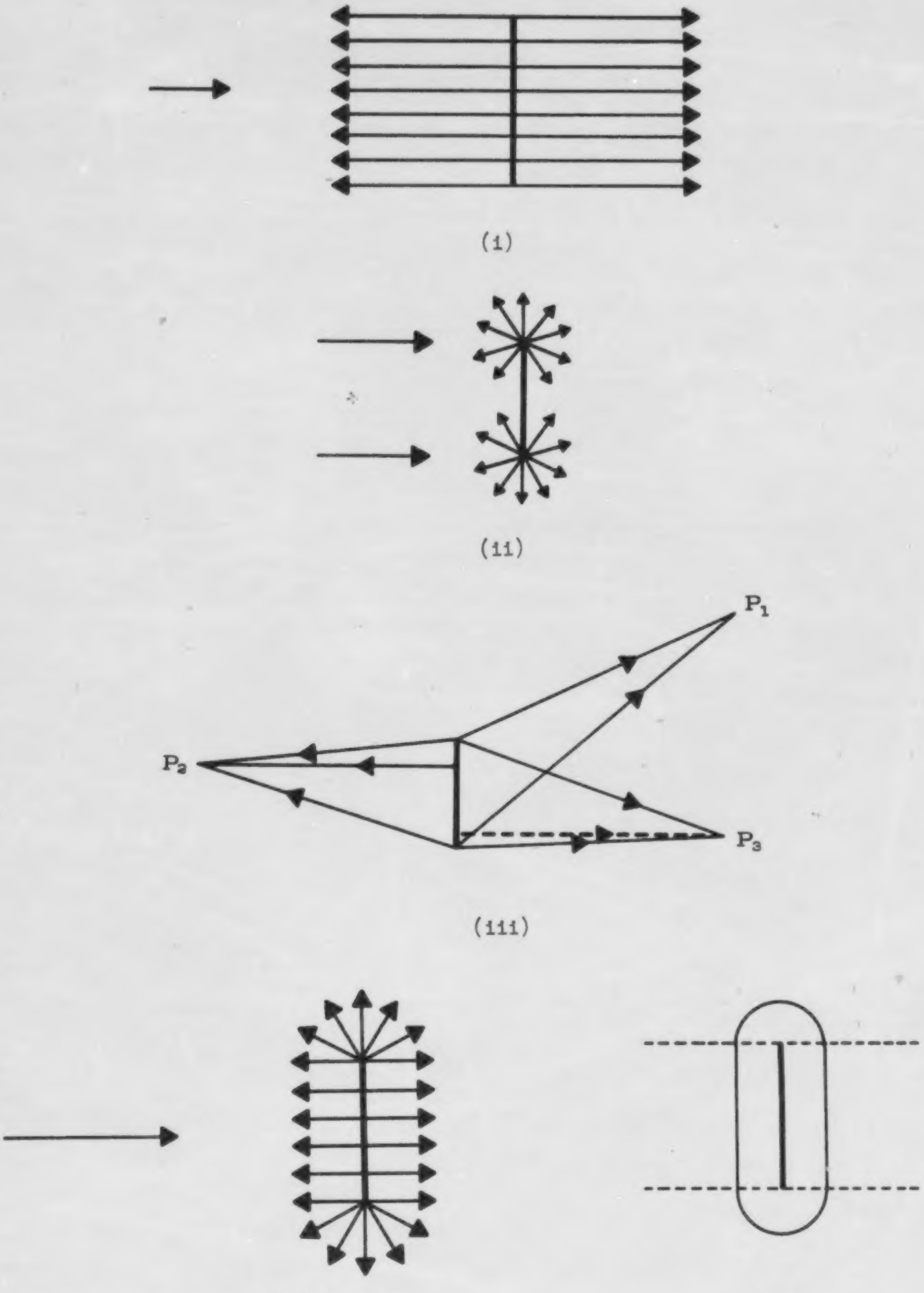


Figure 15-2o

(v) Edge Diffracted Rays.

There are additional sets of rays implicit in Hero's principle, and their utility has been shown by the recent investigations of J.B. Keller. In particular we consider edge diffracted rays arising when a ray is incident on a sharp edge (which "breaks up" or "diffracts" an incident ray). In order to motivate introducing such rays, let us review the preceding material.

We have discussed reflected rays and shadow forming rays, and we saw in connection with the semi-circular cylinder that both kinds of rays were required to obtain a complete coverage of space by scattered rays (or equivalently to obtain closed scattered eikonals). However if the scatterer is a strip as in Figure 15-2p(i), such rays alone do not cover space, which implies that the scatterer's influence is restricted to the two directions shown in the figure. To construct a scattered ray system that covers all space, we introduce the edge diffracted rays of Figure 15-2p(ii); these rays are included in [H] , i.e., an incident ray striking the edge is diffracted to P via the shortest path.



(i)

(ii)

(iii)

(iv)

(v)

Figure 15-2p

From Figure 15-2p(i) and Figure 15-2p(ii), we see that there are essentially three different cases that arise for a fully illuminated strip; these correspond to the three different observation points of Figure 15-2p(iii). An observation point at P_1 receives two diffracted rays; P_2 receives one reflected ray and two diffracted rays; P_3 receives one shadow-forming ray and two diffracted rays. In a subsequent section we show that the magnitude (of energy flow) associated with a diffracted ray is in general much smaller than the magnitude of the other rays in Figure 15-2p(iii). If we assume this result for present purposes, we neglect the diffracted rays in the regions corresponding to P_2 and P_3 and obtain the scattered ray system of Figure 15-2p(iv); this figure shows only the "strongest" scattered ray at each observation point. A corresponding eikonal curve normal to the rays of Figure 15-2p(iv), is shown in Figure 15-2p(v), and it is clear that such surfaces become more circular with increasing distance from the scatterer.

The various rays of Figure 15-2p correspond only to the scattered ray system, i.e., to the effects in space arising from something that obstructs the incident rays; this figure does not take into account that the observation point is also reached by an incident ray. In particular, as discussed for Equations (30) and (31), the incident rays and shadow-forming rays cancel in the shadow region corresponding to P_3. Thus the net effect in the shadow region must arise from the edge diffracted rays as in Figure 15-2q; such effects have been discussed in detail by J.B. Keller. (Bright areas in the shadow region of obstacles with very regular edges were first commented on by Grimaldi, 1613-1663.)

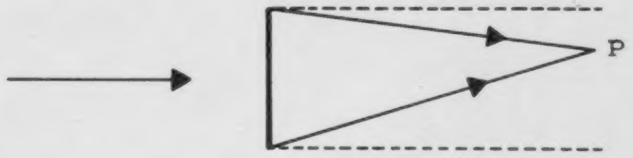


Figure 15-2q

For present purposes, we consider only the caustic of the edge rays for the analogous problem of a disk. Thus if a parallel set of rays is normally incident on a circular disk as in Figure 15-2r, each point of the edge gives rise to a "full fan of rays" normal to the edge at that point. An off-axis observation point receives edge rays only from two points of the circumference on the disk, i.e., from the diametrically opposite points cut by the plane containing the observation point and the disk's axis. However, a point on the

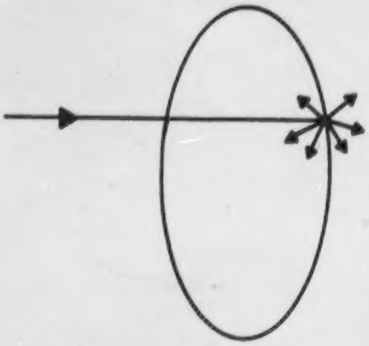


Figure 15-2r

axis of the disk receives edge rays from the entire circumference: the axis is a caustic of the edge rays. Thus the center of the shadow of a normally illuminated circular disk should show a bright spot, the *Arago bright spot*, or *Poisson bright spot* (as predicted originally around 1800 via a wave argument -- a big argument).

For the circular disk, the line caustic of the edge rays is the envelope of the planes normal to the edge of the disk. For a disk of general shape (an arbitrary planar scatterer) normal to the parallel incident rays, the corresponding caustic of the edge rays is a cylindrical surface, the envelope of the planes normal to the edge. Since two such planes intersect in a line normal to the disk, the caustic cylindrical surface generated by the lines of intersection is also normal to the plane of the disk. The cross-section of the caustic cylinder cut by the plane of the disk (or as viewed on a screen in the disk's shadow), is the line envelope of normals to the edge in the plane of the disk; it is the evolute of the edge.

In particular for an elliptic edge

$$(32) \qquad \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \; ,$$

the equation of the evolute is

$$(33) \qquad (ax)^{2/3} + (by)^{2/3} = (a^2 - b^2)^{2/3}$$

(see Exercises 11-6, No. 11). This is the four-cusped curve sketched in Figure 15-2s. Such caustic sections were photographed by Coulson and Becknell in 1922.

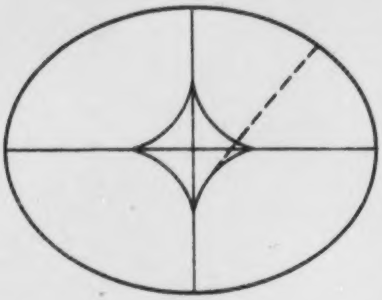


Figure 15-2s

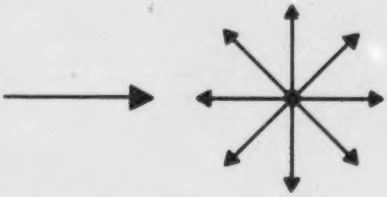


Figure 15-2t

To derive (33) we describe the ellipse parametrically by

(34) $$\xi = a \cos \phi \ , \quad \eta = b \sin \phi \ .$$

The corresponding normal at (ξ, η) through $P = (x,y)$ is specified by

(35) $$g(\phi, P) = \frac{ax}{\cos \phi} - \frac{by}{\sin \phi} - a^2 + b^2 = 0 \ ,$$

and the derivative with respect to ϕ gives

(36) $$D_\phi \, g(\phi, P) = \frac{ax}{\cos^3 \phi} + \frac{by}{\sin^3 \phi} = 0 \ .$$

Substituting (36) in (35) to eliminate either x or y, we see that

(37) $$\frac{ax}{\cos^3 \phi} = \frac{-by}{\sin^3 \phi} = a^2 - b^2 \ .$$

Consequently the locus of the normals is

(38) $$x = \frac{a^2 - b^2}{a} \cos^3 \phi \ , \quad y = \frac{a^2 - b^2}{b} \sin^3 \phi \ .$$

Eliminating ϕ from (38), we obtain the required result (33).

The associated length. In the above we discussed rays reflected from surfaces and rays diffracted by edges. The edge rays as in Figure 15-2p(ii) are drawn radially outward from a point on the line representing the edge, but the scattered rays of Figure 15-2n for the cylinder are not radial. If we visualize the cylinder becoming thinner and thinner we might expect on the basis of our remarks for edge rays that in the small diameter limit the ray system of Figure 15-2n could be represented as a set of radial lines as in Figure 15-2t, but later we shall see that a new phenomenon, diffraction, becomes important in the small diameter limit.

The situations of both Figure 15-2n and Figure 15-2t are covered by [H'], and both correspond to scattering by a circular cylinder. In order to distinguish them we must associate a scale factor for length with light. To do so, in addition to the geometrical property assigned to a ray by [H'], we shall also need to know that light of a single color has an associated length λ (later we shall identify λ as the wavelength), which is independent of the length of the ray path. We could then distinguish the two different scattering situations for the cylindrical obstacle of Figure 15-2n and Figure 15-2t as follows: the ray system of Figure 15-2n corresponds to a very large cylinder* $a \gg \lambda$, and the ray system of Figure 15-2t corresponds to a very small cylinder $a \ll \lambda$.

The existence of an associated length might have been guessed (from Grimaldi's observations of color effects in experiments on light diffracted into shadow regions, Section 15-6) but was not. We show subsequently that the required associated length emerges naturally as part of a more general model, the wave model, for such phenomena. We mention the matter now partly in anticipation, but primarily to stress that the present model is incomplete.

---

*The notation $a \gg \lambda$, (read, "a is much greater than λ) implies that we consider asymptotic expressions in which the parameter $\frac{\lambda}{a}$ is presumed to be small.

Exercises 15-2

1. Show that the shortest path from a point A to a point B by way of a point on a plane mirror, must necessarily lie in the plane containing A and B which is perpendicular to the plane of the mirror.

2. (a) Equation (2) is a necessary but not sufficient condition for the path length L to be a minimum. Show, in fact, that the condition $\alpha = \gamma$ is sufficient for a minimum.

   (b) Show that $\gamma = \alpha$ corresponds to the shortest path by the methods of elementary geometry. (Hint: Use the image principle. This was the method used originally by Hero.)

3. Show that [E2] yields the longest possible reflection path between diametrically opposite points of a circular reflector (Figure 15-2f(i)).

4. Show to a first approximation for a small aperture concave mirror (Figure 15-2f(ii)) that all rays from a source on the axis of the mirror at distance u from the mirror center are reflected through a point at distance v from the mirror, where

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{a} .$$

5. For the semi-circular mirror show that the cusp of the caustic $(\alpha = 0)$ corresponds to $D_\alpha^2\, g(\alpha,P) = 0$.

6. Consider the elliptical reflector,

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 , \qquad (a > b) .$$

   Show that all the rays originating at one focus of the ellipse are reflected through the other focus. (The foci of the ellipse are the points $(\pm\, c,0)$ where $c = \sqrt{a^2 - b^2}$.)

7. Verify analytically that the radius of curvature at a point P of a reflected eikonal for a semi-circular mirror (Figure 15-2ℓ) is $R + \frac{a}{2}\cos\alpha$, where R is the distance from P along the ray to the mirror.

8. Find the edge ray caustic for a parabolic disk with edge given by

$$y^2 = 4px .$$

15-3. Refraction.

In the preceding sections we considered a set of rays incident on reflecting surfaces, and used [H] to determine the reflected set of rays. We now extend the development to partially transparent surfaces and consider in addition a set of transmitted rays. A transmitted ray does not lie in general along the extension of the corresponding incident ray, but makes an appropriate angle with the ray extension; this kind of "break" in the ray path is called refraction.

Observations and studies of the broken appearance of a rod partially immersed in water, and of a beam of light traveling partly in air and partly in water, go back to Euclid and Ptolemy (second century of this era), but the complete description of such effect was first given by Snell (1591-1626). As the appropriate analog of [E] for reflection, we have Snell's Law of Refraction:

> [S] : A ray of light (of one color) incident on the smooth plane interface between two transparent media gives rise (in addition to the reflected ray) to a refracted ray on the other side of the interface. The incident ray, the refracted ray, and the surface normal lie in the same plane, and the two rays are on opposite sides of the normal. The sine of the angle β that the refracted ray makes with the normal is proportional to the sine of the angle α of incidence.

From [S] , we specify the direction of the refracted ray by

$$\mu_2 \sin \beta = \mu_1 \sin \alpha \, , \tag{1}$$

or equivalently by

$$\mu \sin \beta = \sin \alpha \, . \tag{2}$$

The constants μ_1 , and μ_2 are called the indices of refraction, and $\mu = \frac{\mu_2}{\mu_1}$ is called the relative index of refraction. The situation is shown in Figure 15-3a for $\mu_1 < \mu_2$ (this is assumed in all that follows): the ray travels from S to P via a point I on the interface. The μ's are physical constants which specify the essential physical property of the media for the topic at hand; they may be measured experimentally, and we assume they are known. In particular for yellow light passing from air to water, we have $\frac{\mu_1}{\mu_2} \approx \frac{3}{4}$.

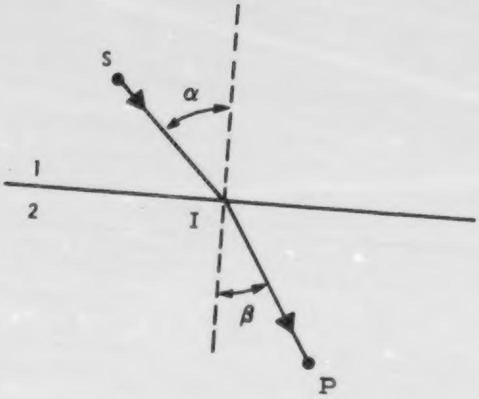


Figure 15-3a

We may apply [S] to such problems as a point source above or below an air-water surface. In particular the caustic for the system of refracted rays can be found by the method of Section 15-2.

Fermat assumed that in a given medium light travels with a velocity v inversely proportional to the index of refraction, $v = \frac{c}{\mu}$ where c is the speed of light in a vacuum and rewrote (1) as

$$\frac{\sin\alpha}{v_1} = \frac{\sin\beta}{v_2}\ ,\quad v_1 = \frac{c}{\mu_1}\ ,\quad v_2 = \frac{c}{\mu_2}\ . \tag{3}$$

He then derived (3) from the following minimum principle called Fermat's Principle.

[F] : A ray takes the least time to travel between two points.

If the total ray path consists of two straight lines L_1 and L_2 in two media with velocities equal to v_1 and v_2 respectively, then the corresponding travel times are $t_i = \frac{L_i}{v_i}$ with $i = 1 , 2$; from [F] we see that $t_1 + t_2$ must be a minimum. Fermat's Principle [F] not only replaces the clumsy [S] (the way [H] replaced [E]), it also includes [H] as the special case where $v_1 = v_2$ and the points S and P are on the same side of the interface.

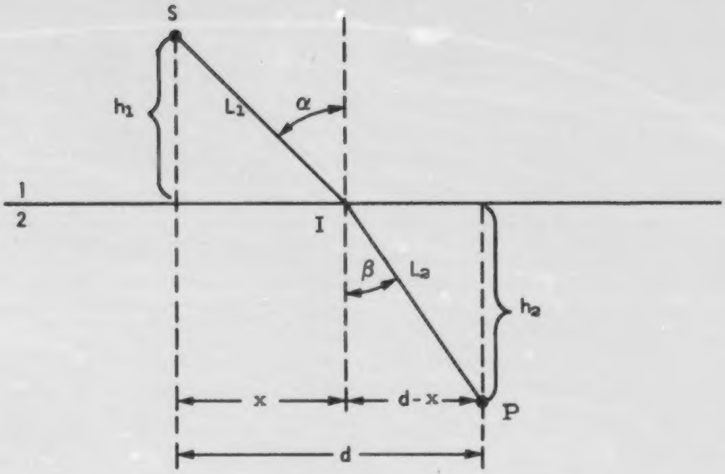


Figure 15-3b

We apply [F] to the configuration of Figure 15-3b to derive [S], essentially as we used [H] to derive [E]. The time taken to go the distance L_1 from S to I at a velocity v_1 is $t_1 = \frac{L_1}{v_1}$, and, similarly, $t_2 = \frac{L_2}{v_2}$ is the travel-time between I and P at velocity v_2 in medium 2. Thus [F] requires that

$$\frac{L_1}{v_1} + \frac{L_2}{v_2} = \frac{\sqrt{h_1^2 + x^2}}{v_1} + \frac{\sqrt{h_2^2 + (d - x)^2}}{v_2} \tag{4}$$

be a minimum. Differentiating (4) with respect to x and equating the result to zero, that is,

$$\frac{x}{v_1\sqrt{h_1^2 + x^2}} - \frac{d - x}{v_2\sqrt{h_1^2 + (d - x)^2}} = \frac{\sin\alpha}{v_1} - \frac{\sin\beta}{v_2} = 0 ,$$

we obtain [S] in the form (1), (2) or (3):

$$\sin\alpha = \frac{v_1}{v_2}\sin\beta = \frac{\mu_2}{\mu_1}\sin\beta = \mu\sin\beta . \tag{5}$$

If $\mu = 1$, and S and P are both in medium 1, then (5) reduces to [E].

It is clear from our discussion of the replacement of [H] by [H'] in Section 2, that we should also generalize [F] by replacing *least* time by *stationary* time. Equivalently, if we define the *optical path length* to be μL, then as the analog of [H'] we take

[F'] : a ray is the stationary optical path between points.

Unlike [S], we may use [F'] and (5) for refraction at curved interfaces. Thus we could now consider the refraction analogs of the reflection problems we considered previously. However, we leave these as exercises and go on to other questions.

Rainbow caustics. Newton (1719) showed that white light could be regarded as composed of different colors, each specified by a different value of some physical parameter (say ω), and that in general the relative index of refraction between two media depended on color, $\mu = \mu(\omega)$. Thus a ray of white light incident at an angle μ on an interface may be treated as a set of coincident rays of different colors (ω) each being refracted a different angle $\beta(\omega)$ as determined by the corresponding index of refraction $\mu(\omega)$. Consequently, a single ray of incident white light becomes a fan of colored rays (the *spectrum*) on refraction, the different colors appearing at angles determined by

$$\sin \beta(\omega) = \frac{\sin \alpha}{\mu(\omega)} . \tag{6}$$

For yellow light incident on an air-water interface we have $\mu = \frac{4}{3}$; for the colors red through yellow on to blue, $\mu(\omega)$ increases through $\frac{4}{3}$ and consequently $\sin \beta(\omega)$ decreases from red to blue as sketched in Figure 15-3c.

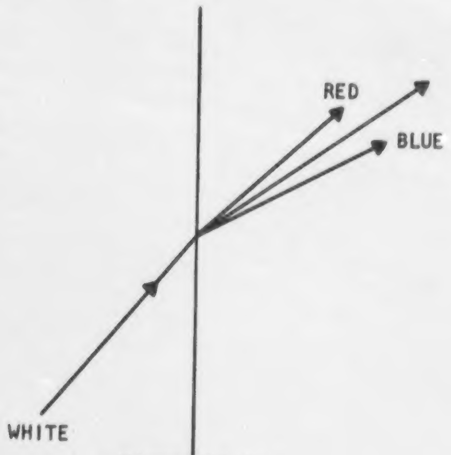


Figure 15-3c

Relation (6) is strikingly exhibited in the rainbow formed by sunlight incident on spherical water drops. In the following we use the methods of calculus to determine the angles of the primary rainbow and secondary rainbow for circular cylinders and spheres.

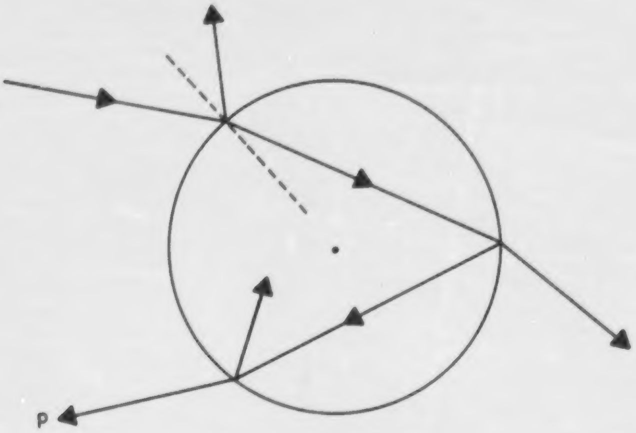


Figure 15-3d

A ray incident on a transparent circle (such as a cylinder of water in air) gives rise to an infinite number of rays. Some of these are shown in Figure 15-3d; initially we consider the ray p. If a system of parallel rays is incident on the cylinder, then we deal with incident rays making all angles α (from 0 to 90°) with the cylinder's normals and to each corresponds a different $p(\alpha)$. We want to show that in the vicinity of some particular value of the angle α (say α_s) the rays $p(\alpha_s)$ will be "focused" (in the sense that they meet at a cusp of the caustic), or equivalently that the angle has a stationary value ϕ_s corresponding to α_s.

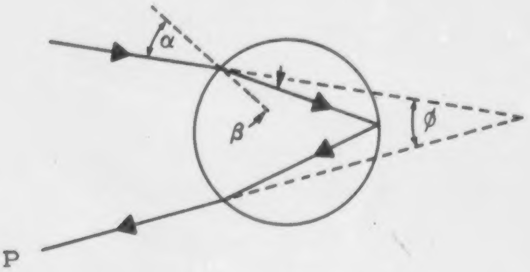


Figure 15-3e

The primary rainbow corresponds to rays that have undergone two refractions and one internal reflection as shown in Figure 15-3e. We now show that the angle ϕ (the angle between the emergent ray and the incident ray) has a stationary value ϕ_s and express ϕ_s in terms of the relative index.

From the figure, we have:

(7) $$\frac{\phi}{2} = 2\beta - \alpha .$$

Equating $\frac{d\phi}{d\alpha}$ to zero we obtain

(8) $$\frac{2d\beta}{d\alpha} = 1 .$$

In addition, from the law of refraction $\mu \sin \beta = \sin \alpha$, we have

(9) $$\mu \cos \beta \frac{d\beta}{d\alpha} = \cos \alpha ,$$

so that (8) and (9) yield

(10) $$\mu \cos \beta = 2 \cos \alpha .$$

Thus from (10) and the law of refraction, we obtain

(11) $$3 \cos^2 \alpha_s = \mu^2 - 1 ,$$

which determines the stationary value α_s of the angle of incidence, and consequently the corresponding values of β_s and ϕ_s . In particular,

(12) $$\sin \frac{\phi_s}{2} = \frac{1}{\mu^2}\left[\frac{4 - \mu^2}{3}\right]^{3/2} .$$

For yellow light, $\mu(\omega) \approx \frac{4}{3}$, and consequently $\phi_s \approx 42^\circ$; for the colors red through blue, the corresponding values of ϕ_s decrease through 42° .

This result for a cylinder also holds for a sphere, and is therefore basic to the rainbow formed when sunlight illuminates a region of air containing many water drops. For one water sphere, if the sun is in back of you and you can see the ray through P of Figure 15-3e, the colored rays will be at about 42° with respect to the direction of incidence (in the plane of the sun, the drop, and your head).

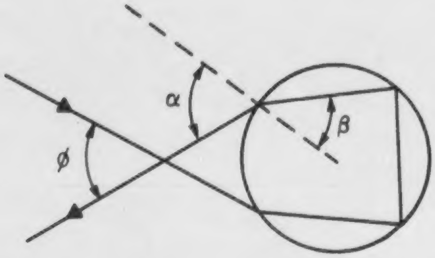


Figure 15-3f

For the *secondary rainbow*, corresponding to two internal reflections, we have the configuration of Figure 15-3f. In this case,

$$\frac{\pi}{2} - \frac{\phi}{2} = 3\beta - \alpha . \tag{13}$$

Differentiating in (13) with respect to α , and using (9) and (5), we obtain

$$8 \cos^2 \alpha_s = \mu^2 - 1 , \tag{14}$$

and consequently

$$\sin \frac{\phi_s}{2} = \frac{\mu^4 + 18\mu^2 - 27}{8\mu^3} .$$

For $\mu = \frac{4}{3}$, we have $\phi_s \approx 51^\circ$. More generally, for n internal reflections, we have

$$\cos^2 \alpha_s = \frac{(\mu^2 - 1)}{n(n + 2)} \tag{16}$$

(see Exercises 15-3, No. 4).

Stratified Medium. If we apply the law of refraction to a ray traveling through a set of parallel slabs as in Figure 15-3g, such that each slab has a different index of refraction, we obtain

$$\mu_0 \sin \theta_0 = \mu_1 \sin \theta_1 = \mu_2 \sin \theta_2 = \ldots = \text{constant} = c . \tag{17}$$

Similarly for the limiting case of a continuum whose index of refraction is solely a function of x , we have

$$\mu(0) \sin \theta(0) = \mu(x) \sin \theta(x) = c . \tag{18}$$

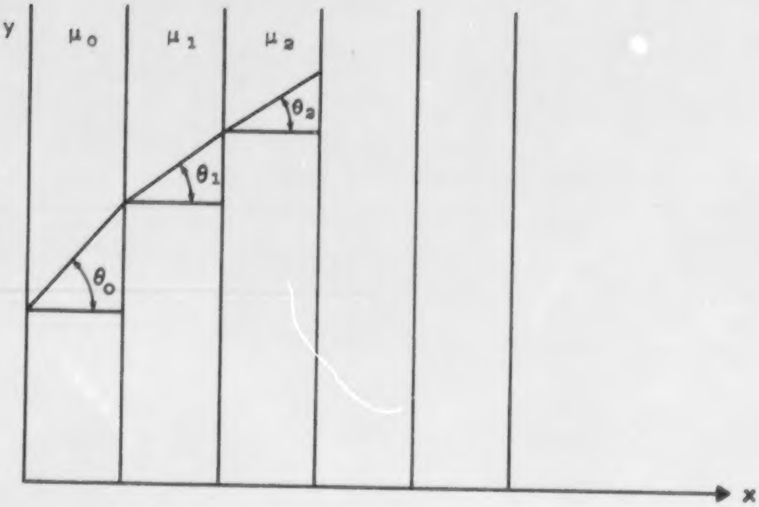


Figure 15-3g

Using $\frac{dy}{dx} = \tan\theta = \frac{\sin\theta}{\sqrt{1 - \sin^2\theta}}$, we have

(19)
$$\frac{dy}{dx} = \frac{(\frac{c}{\mu})}{\sqrt{1 - (\frac{c}{\mu})^2}}$$

from which

(20)
$$\frac{dy}{dx} = \frac{1}{\sqrt{(\frac{\mu}{c})^2 - 1}} .$$

Integrating (20) between 0 and x , we obtain, with $\mu = M(x)$,

(21)
$$y - y_0 = \int_0^x \frac{d\xi}{\sqrt{[\frac{M(\xi)}{c}]^2 - 1}} .$$

Thus in terms of $M(x)$ we have derived an equation to specify the set of rays that start at $(0,y_0)$ and arrive at (x,y) .

As an illustration, we assume

(22)
$$\mu = M(x) = \frac{1}{1 + bx} .$$

To evaluate the integral (21) in terms of (22), we make the substitution

$$c(1 + b\xi) = \sin \phi \ , \tag{23}$$

and rewrite (21) as

$$y - y_0 = \frac{1}{cb} \int_{\sin^{-1} c}^{\sin^{-1} c(1+bx)} \sin \phi \, d\phi \ . \tag{24}$$

Thus on integration, we obtain

$$[x + \frac{1}{b}]^2 + \left[y - y_0 + \frac{\sqrt{1 - c^2}}{cb} \right]^2 = \frac{1}{(cb)^2} \ , \tag{25}$$

i.e., the equation of a circle of radius $\frac{1}{cb}$ whose center is located at

$$\left(- \frac{1}{b} \ , \ y_0 - \frac{\sqrt{1 - c^2}}{cb} \right) \ .$$

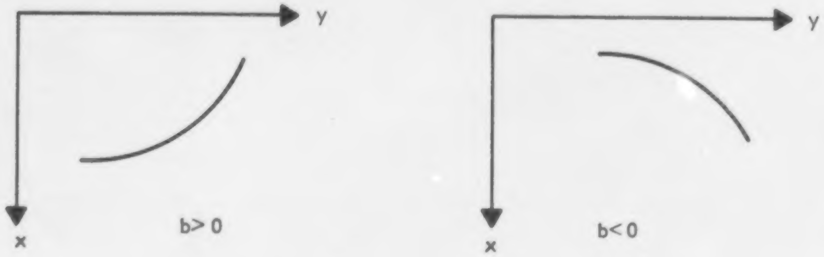


Figure 15-3h

Rotating the coordinate frame of Figure 15-3g (for convenience in the following application to rays in the atmosphere), we show ray paths in Figure 15-3h for (25) with $b > 0$ and with $b < 0$.

The above results serve to account for mirages. Normally the density of the atmosphere decreases gradually with increasing altitude; the index μ , which depends primarily on the density, also decreases gradually. However, over a cold extended surface the density and μ may decrease rapidly with height. An object on the surface may then be seen at large distance by means of downcurving rays as in Figure 15-3i (in which the curvature is greatly exaggerated). The eye sights along the angle of the ray's arrival, and one imagines that the ship lies along the line-of-sight. On a much larger scale and with normal decrease of μ with altitude, Figure 15-3i accounts for our seeing the sun by refraction after it has passed below the horizon.

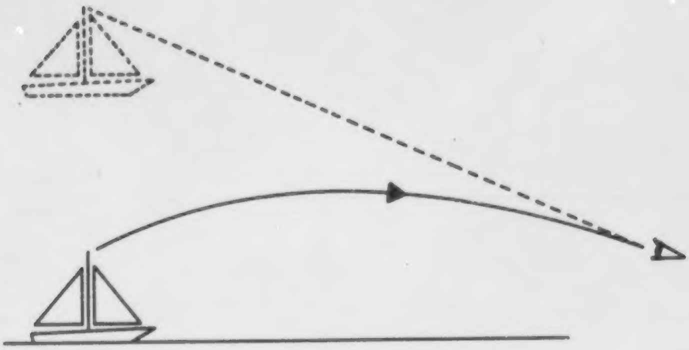


Figure 15-3i

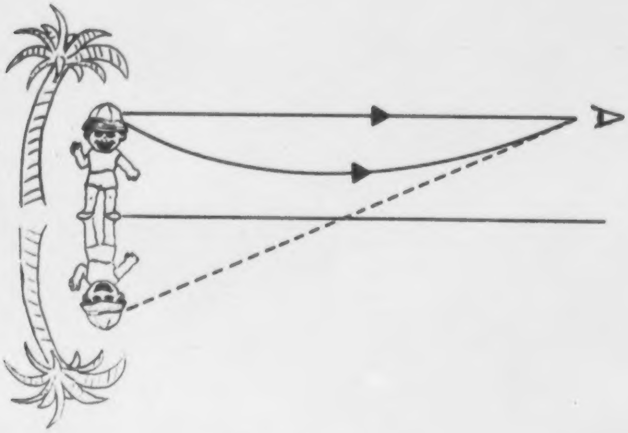
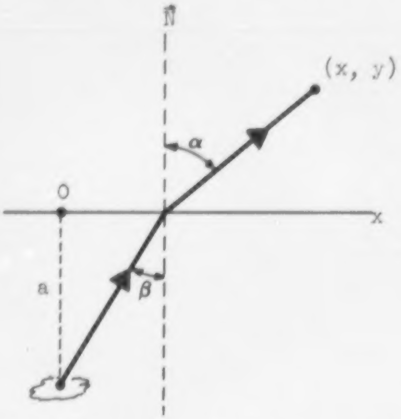


Figure 15-3j

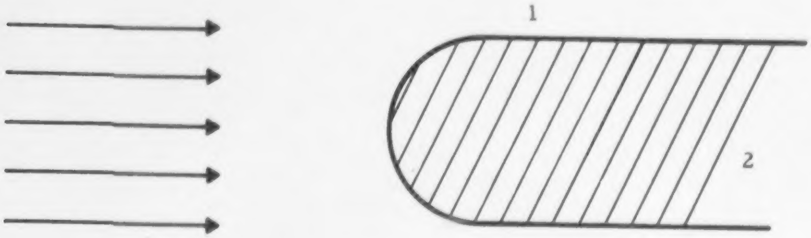
A more common mirage occurs over a hot extended surface when the density and μ first increase and then decrease with increasing height. For such cases the eye may see the object by an upcurving ray as well as by a straight ray as in Figure 15-3j (in which the curvature is again greatly exaggerated). In this situation the eye sees mirror images; since this is reminiscent of reflection on water, one also imagines that a water surface is present.

Exercises 15-3

1. Consider a point source under water $(\mu - \frac{4}{3})$ and the rays for which $\sin\beta < \frac{3}{4}$. Determine the virtual caustic for the rays refracted into air and show that one eikonal is an ellipse. (The apparently different positions of a small pebble in a dish of water as seen from different view points can be described in terms of this caustic.) (Hint: Introduce the parameter $\cos\theta = \frac{\cos\alpha}{\cos\beta}$ where the angles α and β are the angles made with the surface normal in air and water respectively.)



2. (a) Consider the two-dimensional problem of a set of parallel rays in medium 1 incident on a convex semicircle and strip of medium 2 as in the accompanying figure. Obtain the parametric equations for the caustic. Sketch the caustic for $\mu = \frac{4}{3}$.



   (b) Consider the case where medium 2 is a circle. Obtain parametric equations for the caustic of the twice refracted rays. Sketch the caustic for $\mu = \frac{4}{3}$. (Is there a shadow? Try illuminating a cylindrical glass of water with a flashlight.)

3. Consider sunlight illuminating a large number of drops over a very large volume of space and discuss how one will see the familiar arc of the rainbow.

4. Show that ϕ_s is a maximum for the primary rainbow, and a minimum for the secondary bow. Using the fact that $\mu(\omega)$ increases as the colors go from red to blue, state the appearance of the primary and secondary arcs in space and the orders of the colors in the two cases. Derive (16).

5. Sketch the variation of μ with height corresponding to the situations shown in Figure 15-3i and Figure 15-3j.

15-4. *Kepler-Lambert Principle.*

In Section 2, we assumed Hero's principle [H'] that the ray path be stationary, and used the calculus to reveal some of the implicit physics. Except for the discussion of shadows, we did not associate a magnitude with the rays. We now do so, and then supplement [H'] with an *energy principle* or *flux principle*. We introduce a *flux density* F as a measure of the energy flow per second through unit area normal to a ray; indicating the direction of a ray by a unit vector $\vec{R}$, we call $F\vec{R}$ the flux vector.

Kepler in 1604 (by a mixture of mysticism, insight, and some observations of light sources) proposed the *inverse square law* for the flux density associated with a source of light. He argued essentially as follows: If a steady source (one not varying with time) is emitting rays uniformly in all directions, then the total associated flux (total energy per second) passing through any spherical surface centered on the source (as in Figure 15-4a) is a constant; then, since the surface of a sphere increases as the square of its radius r , the flux density F must be proportional $\frac{1}{r^2}$. Equivalently, we may suppose that the flux density depends on radius r alone: $F = \Phi(r)$. Through a small part of the spherical surface of area ΔS the flux is $\Phi(r)\Delta S$. If we subdivide the sphere into small cells and add the contributions from all the cells we obtain a sum analogous to a Riemann sum. We take the limit as the maximum area of the cells approaches zero and call this limit a surface integral, a natural generalization of the concept of Riemann integral. In terms of a surface integral over the sphere the total flux, according to Kepler, is

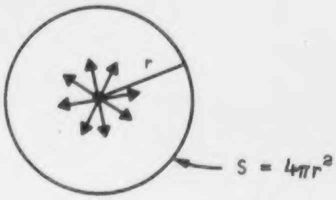


Figure 15-4a

(1) $$\int F\,dS = \int \Phi(r)dS = \Phi(r)\int dS = 4\pi r^2\,\Phi(r)\ .$$

$$= C\ ,\ \text{a constant,}$$

where $\int dS = 4\pi r^2$ is simply the surface area of the sphere. It follows that

(2) $$F = \Phi(r) = \frac{C}{4\pi r^2}\ .$$

Lambert (1760) generalized (1) by taking the component of the flux vector $F\vec{R}$ normal to a surface as the measure of the energy flow. Thus for every surface S enclosing a given steady source (and no other sources), if $\vec{N}$ is the outward unit normal on S , then from the ideas of Kepler and Lambert it follows that:

$$[KL] : \int_S F\,\vec{R}\cdot\vec{N}\,dS = \int_S F\cos\theta\,dS = C \text{ , a constant,}$$

where θ is the angle between the ray direction $\vec{R}$ and the surface normal $\vec{N}$ as in Figure 15-4b.

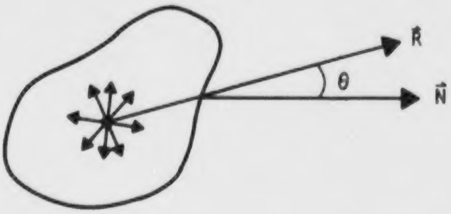


Figure 15-4b

Equation (1) is the special case of [KL] corresponding to a uniform point source at the center of a sphere; for this case F depends only on r , and $\vec{R}$ is parallel to $\vec{N}$. If we take the constant in (1) to equal unity, then the corresponding form of (2) is the flux density for a unit point source:

$$(3) \qquad F = \Phi(r) = \frac{1}{4\pi r^2} ,$$

Equation (3) corresponds to uniform radiation in three-dimensions.

We may also apply [KL] to determine the flux density F for a unit point source radiating uniformly in only two dimensions, or, equivalently, for a unit line source, an extended source along the z-axis which emits rays uniformly in perpendicular xy-planes as in Figure 15-4c. We apply [KL] for $C = 1$, and S equal to a coaxial right circular cylinder having unit length along z and radius r as in Figure 15-4d. The [KL] integral vanishes over the flat caps of the cylinders at $z = \pm\frac{1}{2}$: for these pieces, we see that $\vec{R}$ is perpendicular to $\vec{N} = \pm\vec{k}$ (where k is the unit vector in the direction of the z-axis), and consequently $\vec{R}\cdot\vec{N} = \pm\vec{R}\cdot\vec{k} = 0$. We are thus left with

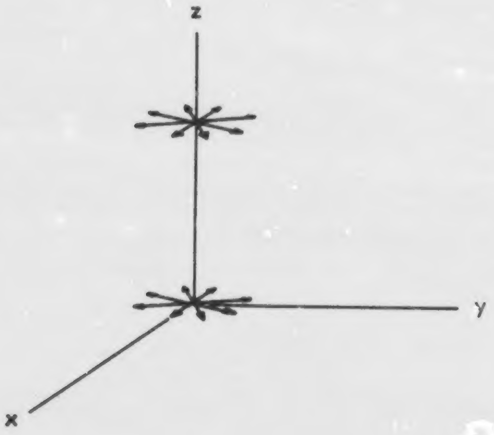


Figure 15-4c

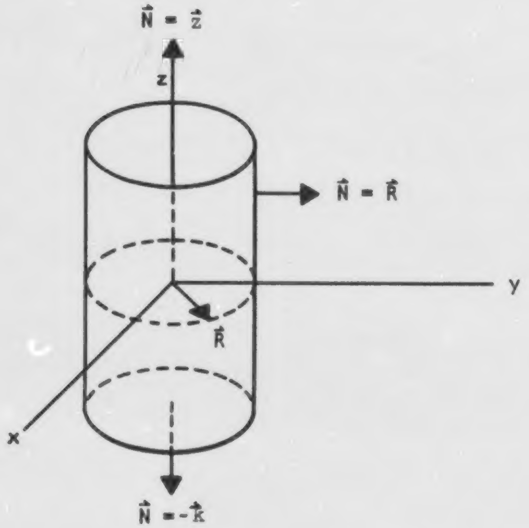

$\vec{N} = \vec{z}$
z
$\vec{N} = \vec{R}$
y
$\vec{R}$
x
$\vec{N} = -\vec{k}$

Figure 15-4d

the integral over the circular wall (of height unity and radius r) for which $\vec{R} \cdot \vec{N} = \vec{R} \cdot \vec{R} = 1$:

(4) $$F \int dS = 2\pi r F = C \ .$$

We take $C = 1$ to define a unit source. Thus from (4),

(5) $$F = \Phi(r) = \frac{1}{2\pi r}$$

is the flux density for unit length of unit line source.

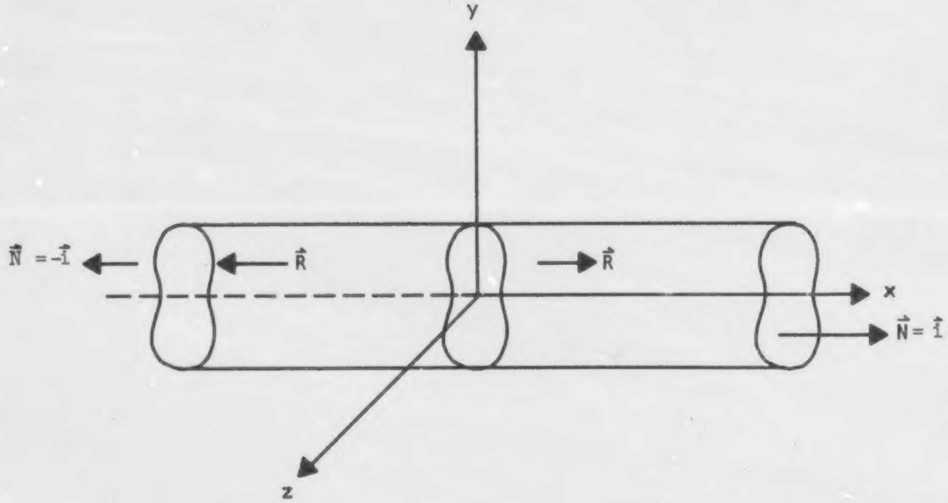


Figure 15-4e

Similarly a planar source is defined as an infinite plane (say zy) emitting rays perpendicularly in the directions $\pm \vec{i}$ where $\vec{i}$ is the unit vector in the direction of the x-axis. For this case we take S as a right cylinder as in Figure 15-4e, with plane faces of unit area parallel to the source (and "enclosing" it). Since $\vec{R}\cdot\vec{N}$ vanishes except over these unit faces, [KL] for $C = 1$ gives

(6) $$F\int dS = F\cdot 2 = 1 ,$$

and consequently

(7) $$F = \frac{1}{2}$$

is the flux density for unit area of source.

It should be kept in mind that all the above equations are very special cases of [KL] . In general $F\vec{R}$ is a function of all coordinates, and [KL] holds for all closed surface enclosing any given set of steady sources.

From [KL] it also follows that the integral over a surface S_0 that does not enclose any sources must vanish:

(8) $$\int_{S_0} F\,\vec{R}\cdot\vec{N}dS = 0 ,$$

i.e., the constant in [KL] is zero for a source-free region. (The source is outside the closed surface, so that whatever flows through part of S_0 flows out through another part.) We use this to define a pencil of rays (a narrow cone of rays) analytically.

Consider the capped tubular surface S_0 of Figure 15-4f which encloses a set of rays. The curved surface S_c is generated by the rays passing through the boundary curve of S_1 and the entrance and exit faces S_1 and S_2 are taken perpendicular to the rays, i.e., the faces are pieces of the eikonal surfaces discussed previously. Thus, if $\vec{N}_c$, $\vec{N}_1$, and $\vec{N}_2$ are the normals to S_c, S_1 and S_2, respectively, then $\vec{R}\cdot\vec{N}_c = 0$, $\vec{R}\cdot\vec{N}_1 = -1$, and $\vec{R}\cdot\vec{N}_2 = 1$. Applying (8) to $S_0 = S_c + S_1 + S_2$, we see that the integral over S_c vanishes and we are left with

(9)
$$\int_{S_1} F dS = \int_{S_2} F dS ,$$

where the integrals are over the entrance and exit faces of the tube. In general F varies from point to point on each face. However, except for special situations, when the faces are small enough so that the variation of F over each is negligible, (9) may be approximated by

(10)
$$F_1 S_1 = F_2 S_2 .$$

The set of rays for which (10) holds is defined as a pencil of rays; the set is inclosed by a tube whose end faces are portions of eikonals. In deriving (10), the "special" situations which are excluded are those for which a face coincides with a focus or caustic. As discussed in Section 15-2, a focus corresponds to the intersection of many rays, so that a closely fitting tube enclosing such a set would narrow down to $S_2 = 0$; for such cases (10) is not a valid relation for F . However, such cases are still covered by (8) provided S_0 does not intersect the caustic.

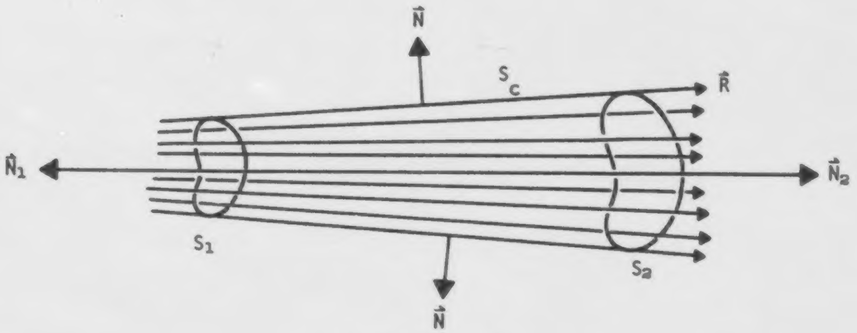


Figure 15-4f

Let us apply (10) to the essentially two-dimensional problem of reflection from a cylindrical surface as described in Section 15-2. We take the z-coordinate in the direction of the generators of the cylinder and put $\Delta S = \Delta z \Delta s$ where s is arclength along a cross-section of the cylindrical surface. Now we drop the unessential z-coordinate and consider the pencil of rays capped by the initial curvilinear element of length Δs_1 and the terminal element of length Δs_2, (Figure 15-4g). To first order we have

(11) $$\Delta s = \rho \Delta \psi$$

where ρ is the radius of curvature at a point of a curvilinear element and $\Delta\psi$ is the angle subtended by the element at the center of curvature 0. Since the two caps are chosen as portions of eikonals (surfaces normal to the rays), and the centers of curvature are the limiting intersection of the common normals to the two caps, we see that the two caps have the same centers of curvature (Figure 15-4g); hence, from (11) $\frac{\Delta s_1}{\rho_1} = \frac{\Delta s_2}{\rho_2}$. Enter this result in (10) to obtain

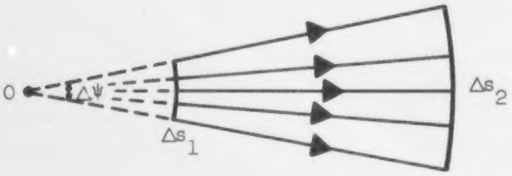


Figure 15-4g

(12) $$F_2 = F_1 \frac{\Delta s_1}{\Delta s_2} = F_1 \frac{\rho_1}{\rho_2} .$$

Equation (12) specifies the variation of the flux density with distance along rays.

Now we consider the perfect (complete reflection of a parallel pencil of rays of width Δs_0 and flux density F_0 from a convex curvilinear portion C_1 of a reflector as in Figure 15-4h. The length of the eikonal of the corresponding reflected pencil is Δs_1 at C_1, and Δs_2 at a distance R from C_1. Perfect reflection means that no rays cross the reflector; i.e., the total incident flux is conserved by the process and passes through the terminal cap Δs_2. Thus (10) holds: $F_0 \Delta s_0 = F_1 \Delta s_1 = F_2 \Delta s_2$. To first order

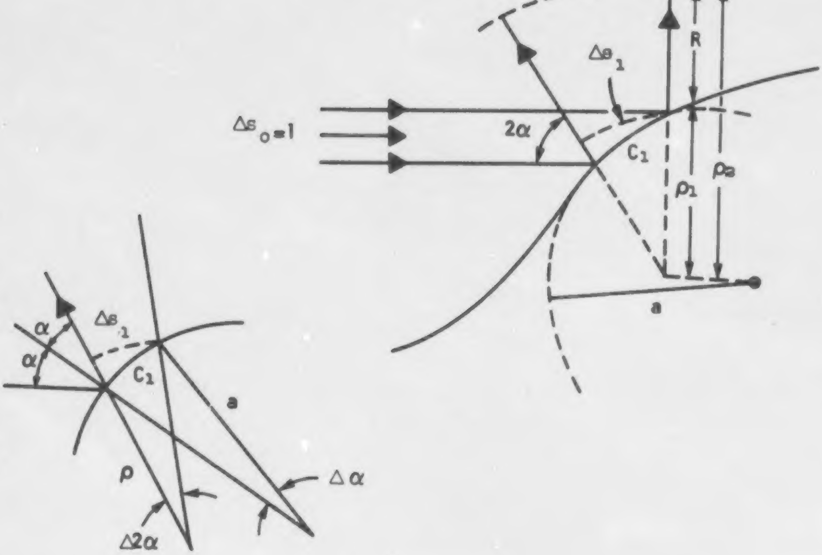


Figure 15-4h

(approximation of the curves by their tangent lines) it is easy to show that $\Delta s_1 = \Delta s_0$; consequently $F_1 = F_0$. Since the reflector is convex, $\Delta s_2 > \Delta s_1$ and it follows that $F_2 < F_1 = F_0$. To determine the exact relation between F_2 and F_0 we now use (12): we write the radii of curvature of the eikonals as $\rho_1 = \rho$ and $\rho_2 = \rho + R$, and obtain

$$F_2 = F_1 \frac{\rho_1}{\rho_2} = \frac{F_0}{\rho + R} , \tag{13}$$

where R is the distance along the reflected rays measured from the reflector, and $\rho + R$ the distance from the caustic (the locus of centers for the radius of curvature the ρ's). Thus, as discussed in Section 15-2, the rays that pass through Δs_2 appear to originate at their virtual intersection point (on the caustic) inside the reflector.

For the semicircular mirror of radius a (see Figure 15-2ℓ(ii)), we found previously that $\rho = \frac{a}{2} \cos \alpha$, where α is the angle of incidence with the surface normal; this also holds to first order for reflection from a convex portion of a more general surface where a is the radius of curvature at the point of incidence. Thus for a convex point (i.e., a point on a convex portion of the mirror), the reflected flux density equals

$$F = \frac{\frac{a}{2}\cos\alpha}{\frac{a}{2}\cos\alpha + R} F_0 \tag{14}$$

(we drop the subscript 2). We may also rewrite (14) as $F = \frac{F_0}{1 + \kappa R}$ where $\kappa = \frac{1}{\rho}$ is the curvature of the eikonal at the reflection point.

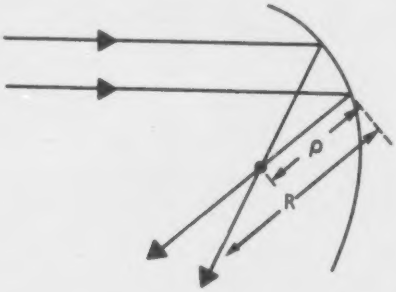


Figure 15-4i

On the other hand, for reflection from a concave point, the caustic is real, and R and ρ are on the same side of the reflector as in Figure 15-4i. For this case we replace ρ by $-\rho$ in (13) and (14), and obtain

$$F = \left|\frac{\rho}{\rho - R}\right| F_0 = \left|\frac{\frac{a}{2}\cos\alpha}{\frac{a}{2}\cos\alpha - R}\right| F_0 \quad , \quad R \neq \rho = \frac{a}{2}\cos\alpha \tag{15}$$

where the absolute value is used because we defined F as a positive quantity. Equation (15) specifies F except on the caustic $R = \rho$, the special situation $(S = 0)$ excluded from the start when we introduced (10).

Flux and Path Length. A remarkable property of F as in (14) (remarkable only at the present stage of our development of a mathematical model for scattering) is that it can be given in the form

$$F = \frac{F_0\, a^2 \cos^2\alpha}{R L_H''} = \frac{F_0 \cos^2\alpha}{R\, D_s^{\,2}\, L_H} \tag{16}$$

where L_H'' is the second derivative with respect to α of the general path length $\xi + R(\theta) = L$ introduced in (3) of Section 15-2, and where the subscript

H indicates that we use the condition $L' = 0$ or $\theta = 2\alpha$, as follows from Hero's principle; D_s^2L is the second derivative with respect to arclength along the reflector. We may assure ourselves that (18) holds by retracing our derivation of the caustic in Section 15-2(iii). Our equation $g(\alpha,P) = 0$ for a reflected ray corresponds to $L' = 0$, and our equation $D_\alpha g(\alpha,P) = 0$ for the caustic of the reflected rays corresponds to $L_H'' = 0$. (See Exercises 15-4, No. 1.)

We mention this now to make explicit that F becomes singular on the caustic $L_H'' = D_\alpha g(\alpha,P) = 0$, which indicates a limitation of our present essentially geometrical model for the propagation of light. Later we shall see this result as a limiting case of a deeper relation between flux and path length that must hold for a more complete model.

Partially transparent surface. We may extend the present flux considerations to the case of partially transparent media considered in Section 15-3, and obtain the corresponding reflected and transmitted fluxes when a pencil of rays is incident on the curved interface of two different optical media. At the present primitive stage of our model, we would simply introduce a reflection factor $0 < P(\alpha) < 1$ as a multiplier of the incident flux to obtain the values of the reflected flux for the corresponding perfectly reflecting surface. Applying (8) to a pencil of rays incident on a plane interface (with S_0 "enclosing" the interface as in Figure 15-4e), we then find that for the incident flux to equal the sum of that reflected and that transmitted we require that the geometrically transmitted flux be multiplied by the transmission factor $1 - P(\alpha)$.

Scattering Applications. In the above we applied [KL] to obtain the flux density for elementary sources (in one-, two-, and three-dimensions), and to determine the change of flux density along a ray in a pencil of varying cross section. We now extend our considerations for the elementary sources to analogous scattering problems. We define the corresponding "elementary scatterers" by the previous stipulation that the total radiated flux equal unity and that it be distributed uniformly over the available directions; then we indicate generalizations. We do not solve any scattering problems explicitly, but exploit the previous development to introduce terms and general forms for subsequent use.

Thus if we have a set of parallel rays normally incident upon a perfectly reflecting planar scatterer at $x = 0$, as in Figure 15-4j, then from Section 15-2, the incident set of rays gives rise to a reflected set of rays and to a

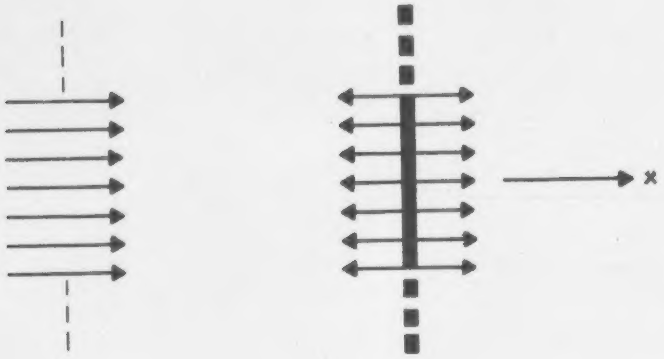


Figure 15-4j

shadow-forming set of rays. We may say that the incident rays have "excited" the plane and converted it to a source of radiation; we call the incident set the primary radiation and the scattered set (reflected plus shadow forming) the secondary radiation, and say that the plane has become a secondary source. We define an elementary planar scatterer as a secondary source fully analogous to the planar source considered in Figure 15-4e and Equations (6) and (7). (In a later section we consider analytically the specific problem to which this corresponds.) The essential feature of (7) is that the flux does not depend on distance. Similarly for a planar scatterer which both reflects and refracts we write the scattered flux corresponding to the direction of incidence $\vec{i}$ parallel to the x-axis, as

$$F = M(\vec{R}) \quad , \quad \vec{R} = \pm \vec{i} \ , \tag{17}$$

where the direction of scattering $\vec{R}$ corresponds either to geometrical reflection, $\vec{R} = -\vec{i}$, or to forward scattering, $\vec{R} = \vec{i}$. (For a perfect reflector, it turns out that M is the square of E_s discussed in Section 15-2(iv): if the incident flux density is unity, then $M = 1$.)

Similarly if we visualize rays incident perpendicularly on a fine cylinder as in Figure 15-4k and apply [H] essentially as for the discussion of edge

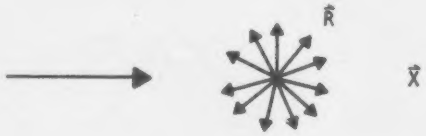


Figure 15-4k

diffracted rays in Section 15-2, we see that the scattered set of rays travel radially outward from the scatterer. We define an elementary line scatterer as a secondary source fully analogous to the line source of Figure 15-4c and Figure 15-4d and Equations (4) and (5), i.e., the total outgoing flux per unit length of scatterer is unity, and the scattered flux density per unit length is given by (5). Similarly for a more general line-like or cylindrical obstruction F is inversely proportional to R but the flux density is no longer the same in all directions:

$$F = \frac{M(\vec{R})}{R}, \tag{18}$$

where the direction of observation $\vec{R}$ may range over all values in the xy-plane.

Finally in three dimensions, we visualize a point scatterer excited by rays, and define a secondary point source analogous to that of Figure 15-4b and Equations (1) to (3). More generally, for an arbitrary scatterer in three dimensions, the analog of (18) is

$$F = \frac{M(\vec{R})}{R^2}, \tag{19}$$

where again M depends only on directions and not on distance. The functions M for the three types of scatterers depend on various parameters, and their determination requires a more complete mathematical model than the present one. However, the forms (17), (18), and (19) give the appropriate dependence of F on R.

We are now in a position to further our discussion of the relative magnitudes of the different rays of Section 15-2. Thus the parallel rays incident on the broad finite strip of Figure 15-2p(i), excites essentially two kinds of secondary sources: the body of the strip becomes a one-dimensional plane-source with reflected flux density equal to that incident, and the edges become secondary line sources with flux density specified by (18). The flux density of the rays geometrically reflected from a plane are independent of distance, but the flux density of the rays diffracted from the edges decreases as $\frac{1}{R}$ with increasing R. Thus in the region of space covered by the reflected rays, the edge rays become relatively weaker with increasing R.

Exercises 15-4

1. Derive the relation between flux and path length, Equation (16), from Equation (14).

15-5. Huyghens' Principle.

All our preceding discussion is covered by the two "laws of nature" [F'] and [KL] plus some of the implicit physics relevant to geometrical optical phenomena. The basic physics was contained in the two laws, the rest was mathematical manipulation based on a geometry of rays and some procedures of the calculus. As a preliminary to the introduction of additional structure into our mathematical model for the propagation of light, we now supplement our previous geometrical construction of the eikonals by an alternative construction called Huyghens' principle. This principle by itself does not give us any new results,but (and this is often much more significant) it gives us a new way of thinking about the results we have already obtained.

There are two familiar forms in which energy propagates: packaged around particles, or associated with waves (e.g., if you are swimming with a friend you may transmit energy to him by splashing and showering him with water drops, or by setting up a wave on the water's surface). At the present stage of the development the flux involved in [KL] is in some sense guided along the geometrical rays. It is easy to visualize the rays as guide lines for very fine particles (a view held by the ancients, and refined by Newton - 1705), but we may also regard the rays as the normals of a system of wave surfaces (the eikonals).

Many individuals (Hooke, Euler, and others) regarded light as a wave motion in a special medium, but it was Huyghens (1690) who introduced the subject as an analytical one. His intuition was based on the analogous two-dimensional problem of how disturbances travel on the surface of water. (Touch the surface of still water and the disturbance travels outward in a circular ripple along the water surface.)

Huyghens used the fact that light has a finite velocity of propagation v (as established experimentally by Romer, 1676) for the development of a wave theory of light. He assumed that in a given medium, light starting from an elementary source at time t_0 would spread as a spherical surface whose radius $r(t)$ increased in time as $v(t - t_0)$.

If we start a light source at time t_0 and leave it on, the corresponding Huyghens' wave surface is an outgoing spherical front -- a step function disturbance whose one-dimensional analog is shown in Figure 15-5a. In this figure we plot a magnitude associated with the disturbance (say the flux density F introduced in Section 15-4, or a related quantity) as a function of time; at time $t_1 > t_0$, the wave front has moved a distance $v(t_1 - t_0)$, and it keeps

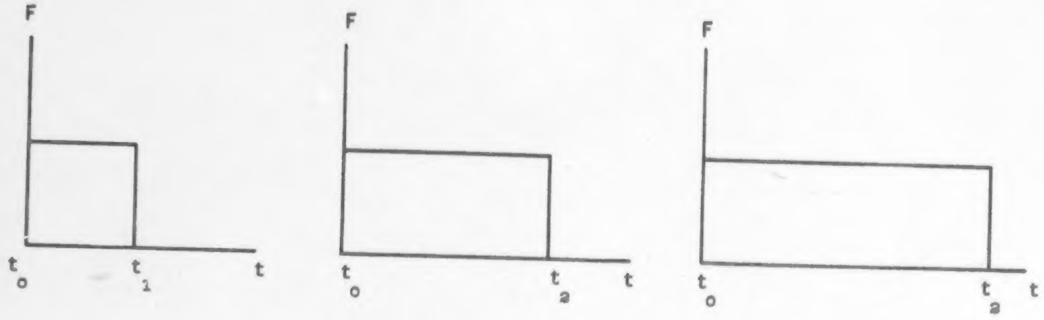


Figure 15-5a

advancing with increasing t . (The discontinuous function drawn in Figure 2 is called a Heaviside pulse.)

Starting with an advancing wave front (hereafter, the wave surface W) in three dimensions, Huyghens regarded each point on the wave surface W as a new source of an elementary spherical wave (call it a wavelet, w) whose radius also increases in time proportionally to v . Thus if the original wave surface W is a sphere of radius $r(t_1)$, the wavelet surface w spreads as a sphere of radius $R(t) = v(t - t_1)$; the two-dimensional analog is shown

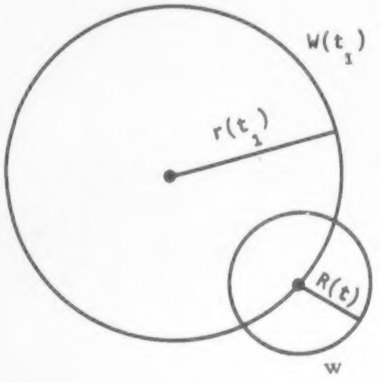


Figure 15-5b

in Figure 15-5b. To obtain the wave surface of the source's advancing wave front, Huyghens prescribed

> [Hu] : to construct the wave surface $W(t_2)$ at time $t_2 > t_1$, regard the wave surface $W(t_1)$ at time t_1 as the locus of the centers of wavelets w of identical radius $R = v(t_2 - t_1)$, and take $W(t_2)$ as the outer envelope of the set of w's .

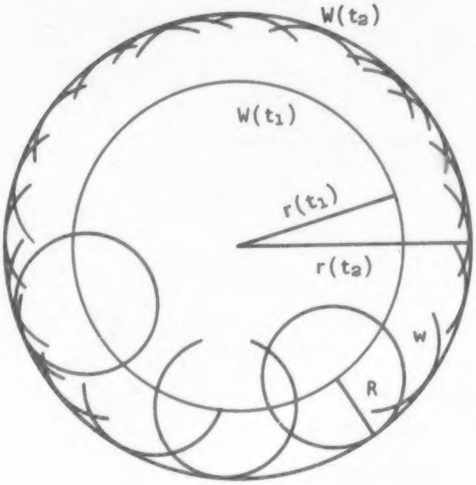


Figure 15-5c

Figure 15-5c, based on Figure 15-5b, illustrates [Hu] . The essential notion is that if we assign a magnitude to the outward part of a wavelet, then only on the outward envelope of the set of w's (i.e., only on $W(t_2)$) do the magnitudes of the w's add up (reinforce) to give a significant overall effect.

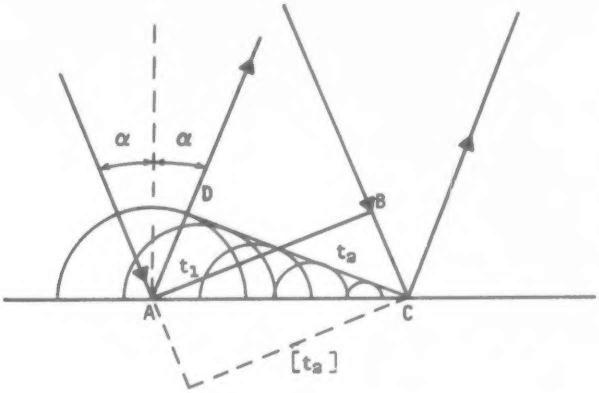


Figure 15-5d

If a planar portion of a wave surface is incident on a reflecting surface, we can construct the reflected wave front by means of [Hu] as indicated in Figure 15-5d. The figure shows the incident wave front at t_1 (plus the two rays or normals that bound it), and a front (shown dashed) at $[t_2]$ to indicate where the incident front would have reached at time $t_2 > t_1$ in the absence of the reflector. The actual reflected front at time t (the image shown unbroken of the dashed front at $[t_2]$) is the envelope of the wavelets generated by the incident front at different times as the point where it encountered the reflecting surface moved off to the right. The dashed front is also the wave front of the shadow-forming rays discussed in Section 15-2. Figure 15-5e shows how Huyghens' construction for scattering by a strip yields the closed scattered wave surface corresponding to the reflected plus shadow-forming plus diffracted rays of Figure 15-2p(iv); the result is of course simply the closed eikonal of Figure 15-2p(v).

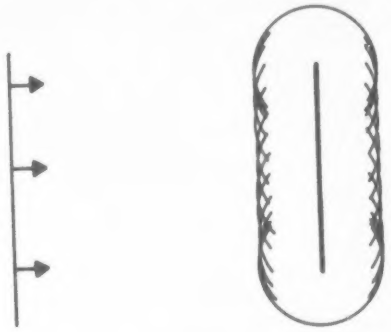


Figure 15-5e

Similarly if the scattering surface is the interface between two different optical media specified by velocities v and v_1, we construct the transmitted portions of the wavelets to take into account that these portions are traveling at velocity v_1 instead of v, and then construct their envelope to obtain the refracted wave front as in Figure 15-5f.

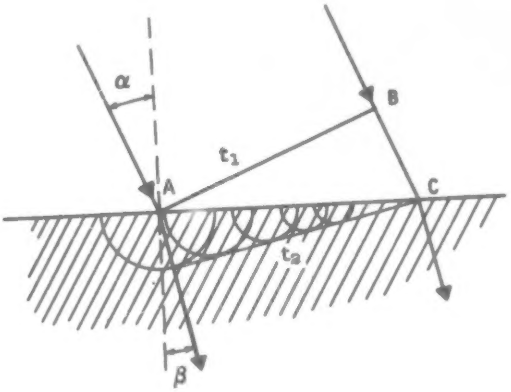


Figure 15-5f

We have indicated that the Huyghens's wave surfaces are simply the eikonal surfaces discussed in Section 15-2. We now apply [Hu] to the reflection of a plane wave front (parallel ray system) by a perfectly reflecting convex semicircle and make this identification explicit. Since all waves in this problem move with the same velocity, all distances (L) traveled are porportional to time (t) , so that we may work with either L or t ; in order to exploit our previous figures and results, we work with distance L . The center of the circle in Figure 15-5g is at $x = 0$, $y = 0$. The corresponding incident wave is a plane wave front whose position at any time $t > t_0 = 0$ may be indicated by $x > x_0 = 0$; i.e., our reference time is $t = 0$, and our reference position is $x = 0$.

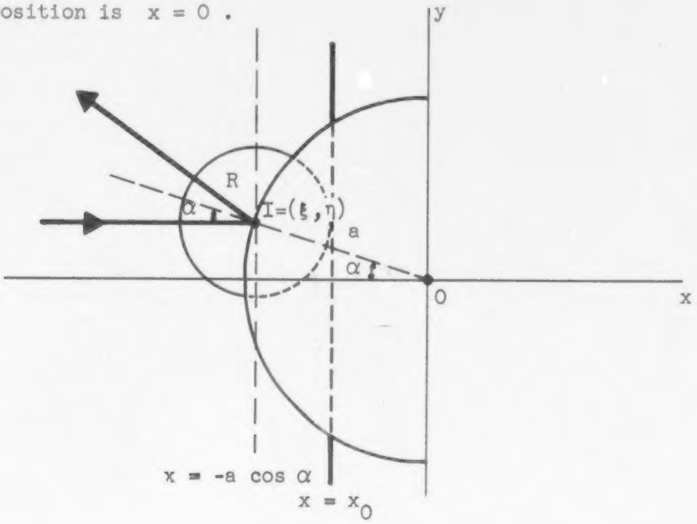


Figure 15-5g

We treat the Huyghens construction for Figure 15-5g essentially as we did that of Figure 15-5d. We construct wavelets of different radii at different points on the circular scatterer, the radius at a point being proportional to the time it would have taken the incident wave front to travel from that point to the plane $x = x_0$. Using Huyghens's principle in this manner we may say that a point $I = (\xi,\eta)$ of the scatterer, where $\xi = -a \cos \alpha < x_0$ under excitation by the wave front $x = -a \cos \alpha$ (see Figure 15-5g) radiates a circular wavelet of radius $|\xi - x_0| = |a \cos \alpha + x_0|$; here $x = x_0$ is the present position of the incident wave front. The resultant wave front is the envelope of all such elementary wavelets. To draw a wave front construct enough such wavelets to enable their envelope to be sketched. Figure 15-5h shows the case $x_0 = 0$ (i.e., for the time when the incident front is at the origin) and Figure 15-2o of Section 15-2 shows additional curves for different values r ; in each case, the straight portion of the curve corresponding to the shadow wave front is also the position that would have been reached by the incident front in the absence of the scatterer. The curves of Figure 15-2o

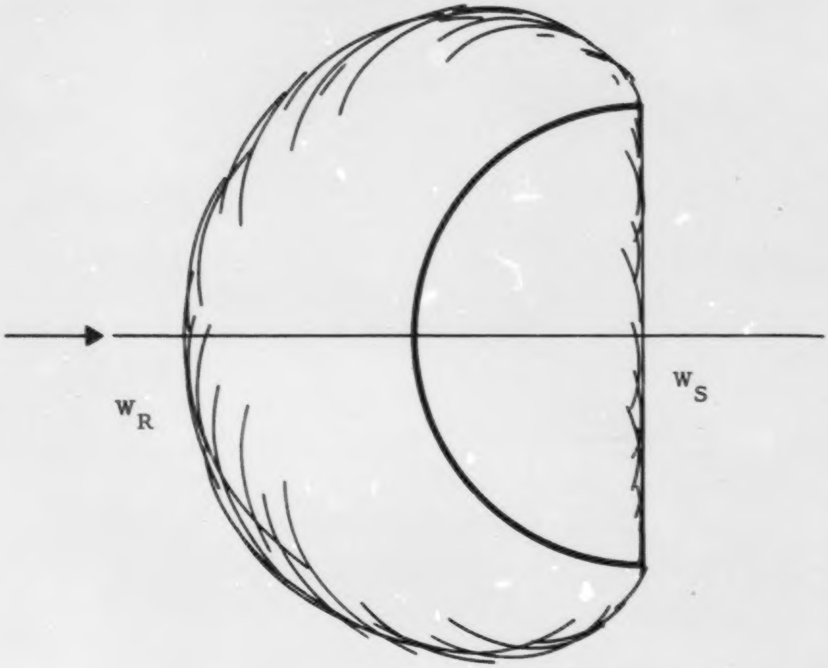


Figure 15-5h

can be constructed either by using the present procedure (circles centered on the scatterer) for different constants, or by using the wave surface of Figure 15-5h as the locus of circles of identical radii and then drawing their outward envelope.

Analytically, we find the envelope of the family of circles by the same procedure we used in Section 15-2 to obtain the envelope of a set of straight lines. Thus if we take $x_0 = 0$, we have the equation of a Huyghens circlet

$$(x + a \cos \alpha)^2 + (y - a \sin \alpha)^2 = a^2 \cos^2 \alpha \ . \tag{1}$$

The derivative with respect to α gives $y = a \sin \alpha - x \tan \alpha$, and entering this expression for y in the equation of the circle (1) gives $x[x(1 + \tan^2 \alpha) + 2a \cos \alpha] = 0$. Thus either

$$x = -2a \cos^3 \alpha \quad \text{and} \quad y = a \sin \alpha(1 + 2 \cos^2 \alpha) \ , \ (-\tfrac{\pi}{2} \le \alpha \le \tfrac{\pi}{2}) \ , \tag{2}$$

or

$$x = 0 \quad \text{and} \quad y = a \sin \alpha \ , \qquad (-\tfrac{\pi}{2} \le \alpha \le \tfrac{\pi}{2}) \ . \tag{3}$$

The parametric Equations (2) describe the envelope

$$\frac{(x^2 + y^2)}{a^2} = 1 + 3(\frac{x}{a})^{2/3} \tag{4}$$

so that the corresponding curve (W_R of Figure 15-5h) is half of a two-cusped epicycloid (twice the size and rotated through 90 degrees, as compared with that for the rays shown in Figure 15-2ℓ(i)). This portion of the wave front is generated by a point on a circle of radius $\frac{a}{2}$ rolling on the circle of radius a. The Equations (3) specify the W_S portion of the envelope of Figure 15-5h, which consists of a line segment of width $2a$ normal to the direction of incidence; this corresponds to the shadow forming wave.

Given the analytical expression for the wave surface analytically in (3) and (4), or its graph, as in Figure 15-5h, we can construct its normals (the rays of Section 15-2), and then obtain any other wave front on laying off a constant distance along the normals and joining the points. We can construct the evolute of the wave fronts (the caustic of the rays) and determine that $R + \frac{a}{2} \cos \alpha$ is the radius of curvature, where we recall that R is distance along the ray from the mirror; and, of course, we can "discover" the law of geometrical reflection by noting that at a given point, the reflected and incident wave normals make equal and opposite angles with the scatterer's normal.

From a "pure" wave view, in order to determine the scattered wave front when the incident front is at any $x_0 > -a$, we use the wave surface of (4) and (3) derived for $x_0 = 0$ ($W = W_R + W_S$ of Figure 15-5h) as the locus of the centers of circles of radius $|x_0|$, and again determine the envelope mechanically or analytically; i.e., we need not refer back to the surface of the scatterer. Thus we determine both past and future wave fronts by the Huyghens construction. If $x_0 > -\frac{a}{2}$ we obtain the wave fronts shown in Figure 15-2o of Section 15-2. The point on a wave surface (corresponding to the incident front at x_0) at a distance $R = x_0 - a \cos \alpha$ along a ray, may also be designated by the cylindrical coordinates r and θ as in Figure 15-5i. For very large values of x_0, we see that R and r are practically parallel and that $\theta \approx \pi - 2\alpha$; we have

$r \approx R - a \cos \alpha = x_0 - 2a \cos \alpha \approx x_0 - 2a\left|\sin \frac{\theta}{2}\right|$, which corresponds to a wave from a source at $x = -\frac{a}{2}$ (the cusp of the virtual caustic). If we then neglect $a \ll r$, we obtain $r \sim x_0$ and the wave fronts approach circles centered on the origin of the mirror.

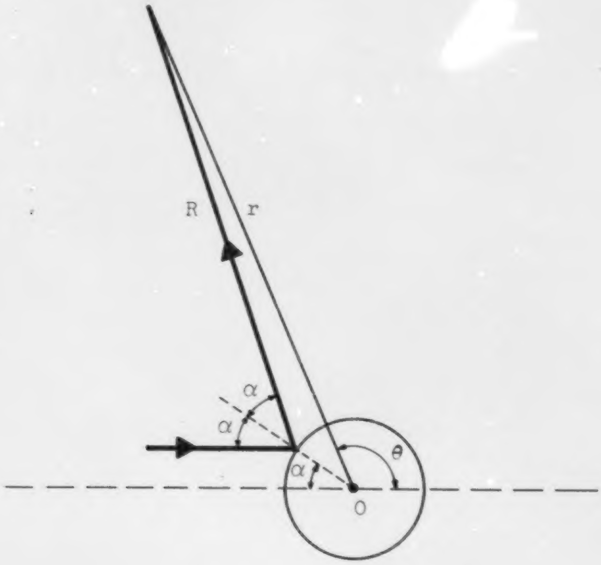
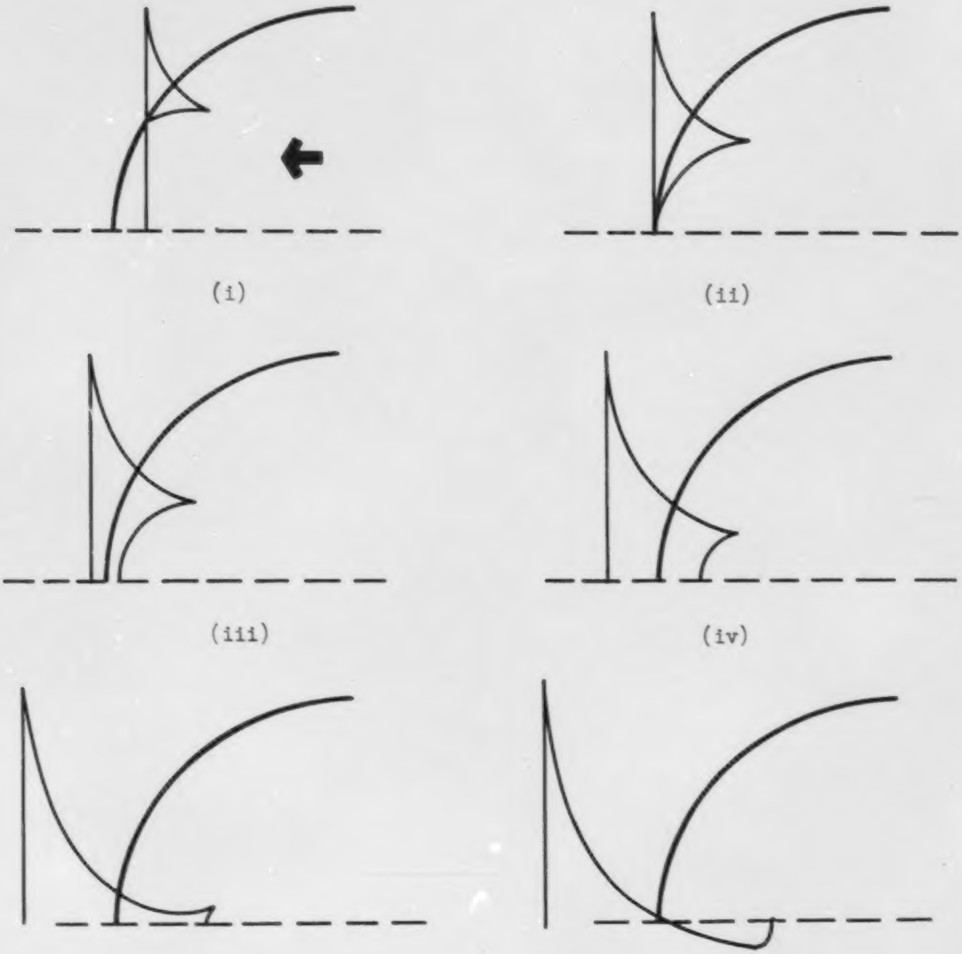


Figure 15-5i

For incidence on the convex semicylinder, these wave fronts are real in the same sense as we spoke of real intersections for the rays of Section 15-2; for incidence on the concave semicylinder, the wave fronts of Figure 15-2o are virtual. The virtual wave fronts for incidence on the convex cylinder (the real ones for the concave case) are obtained for $x_0 < -\frac{a}{2}$; these are the curves shown in Figure 15-5j (i-vi) plus their images in the x-axis. For the sphere we obtain the virtual wave fronts by rotating the curves of Figures 15-5j(i-vi) around the x-axis.

The "inside fronts" nearest the scatterer in Figure 15-5i may be likened to a volcano with crater. The concave craters of successive fronts represent a bounded portion of a wave outgoing from an origin at $x = -\frac{a}{2}$ (the geometric focus). The outer sides of the volcanos correspond to bounded portions of waves outgoing from the edge of the scatterer. The flat bases of the volcanos represent the shadow-forming wave. The most significant feature of the set of figures is that the locus of the cusps is the virtual caustic of the geometrical rays derived previously in Figure 15-2ℓ. Thus in the same



(i) (ii)

(iii) (iv)

(v) (vi)

Figure 15-5j

fashion as we traced the origin of the real rays to their virtual caustic, the real system of wave fronts may be traced back to the cusps of the virtual wave system.

15-6. (i) Periodic Waves.

In the preceding section we obtained the eikonals of ray theory directly by applying [Hu] to a wave front. Huygens represented light essentially as an irregular sequence of isolated disturbances or pulses. The essential feature of the mathematical description of a wave pulse is shown in Figure 15-6a, which represents a disturbance propagating with velocity v along the x-axis. The significant aspect of Figure 15-6a is that the shape of the pulse does not change in time.

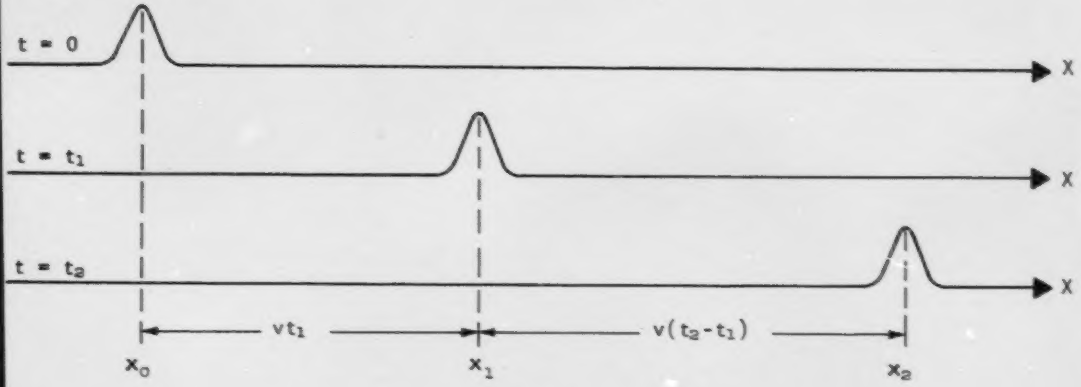


Figure 15-6a

If we specify the pulse form at $t = 0$ by $y = f(x)$, then since the pulse form at any time t is obtained by the translation $x \to x + vt$, the pulse form at time t is given by

$$y = f(x - vt) . \tag{1}$$

Physically, we see that the function $f(x - vt)$ represents the unchanged disturbance moving along the x-axis (direction $\vec{i}$) with constant velocity v. Similarly a disturbance moving in the direction $-\vec{i}$ would be represented by $f(x + vt)$.

By itself [Hu] is merely another method for rederiving the results we obtained geometrically. However if we associate the idea of periodicity with Huygens' idea of waves, then we will have progressed quite far towards the full mathematical model we are developing.

Periodicity. Newton (1642-1727), by refracting a pencil of white light through a prism of glass, showed that a ray of white light could be regarded as made up of rays each having a single color (an idealization called monochromatic light), and that the relative index of refraction μ depends on color. We touched on this before in our discussion of the rainbow when we we worked with $\mu(\omega)$, with ω as the "color parameter." His studies on the colors obtained by illuminating thin transparent plates, essentially established that

[N] : monochromatic light is periodic with period dependent on ω .

Newton's picture of light as a stream of fine particles subject to periodic "fits" that followed each other at regular intervals is not appropriate for the visible phenomena he was familiar with, but the idea of periodicity related to color is as significant as Huyghens's idea of waves.

Young (1801) combined Newton's idea of periodicity with Huyghens's idea of waves, and regarded monochromatic light as made up of periodic waves.

If we rewrite (1) in the form $y = f(p)$ with

$$(2) \qquad p = k(x - vt) \quad , \quad k = k(\omega)$$

where $k(\omega)$ (the "propagation constant") depends on color, and where p is called the phase of the wave, then Young's principle states

[Y] : monochromatic light can be represented by a wave which is a periodic function of the phase $p = k(x - vt)$.

Analytically, we express [Y] in the form

$$(3) \qquad f(p) = f(2\pi + p) = f(2\pi n + p) \; ; \qquad n = 0 \; , \pm 1 \; , \pm 2 \; , \dots ,$$

where the period of f is fixed as 2π ; the term p represents phase or position within a cycle or interval of length 2π .

If we impose the condition

$$(4) \qquad f(0) = A \; ,$$

where A the amplitude is the maximum value of $|f|$, then the simplest wave function satisfying (3) and (4) is the circular function

$$(5) \qquad y = f(p) = A \cos p = A \cos(k[x - vt]) \equiv u(x,t) \; .$$

We may write

$$(6) \qquad k = \frac{2\pi}{\lambda}$$

where λ is the wavelength associated with light of a single color. If we increase x by Δx , then we increase p by $\Delta p = 2\pi \frac{\Delta x}{\lambda}$. Each time x changes by the length λ , we have $\frac{\Delta x}{\lambda} = 1$ and $\Delta p = 2\pi$, and $f(p)$ of (5) goes through a maximum and minimum. The factor $kx = \frac{2\pi x}{\lambda}$ is a convenient dimensionless measure of distance for a monochromatic wave; it gives directly the phase change in units of 2π corresponding to the distance x . Similarly, we may write $kvt = 2\pi \frac{t}{T}$ with

$$kv = \frac{2\pi}{T} \tag{7}$$

as a dimensionless measure of time corresponding to the phase change in units of 2π for a time interval t. From (6) we have $kv = \frac{2\pi v}{\lambda}$ which together with (7) gives

$$v = \frac{\lambda}{T} . \tag{8}$$

We are now in a position to interpret the parameters λ and $T = v\lambda$ introduced in the above as well as the corresponding "color parameter" we have mentioned previously.

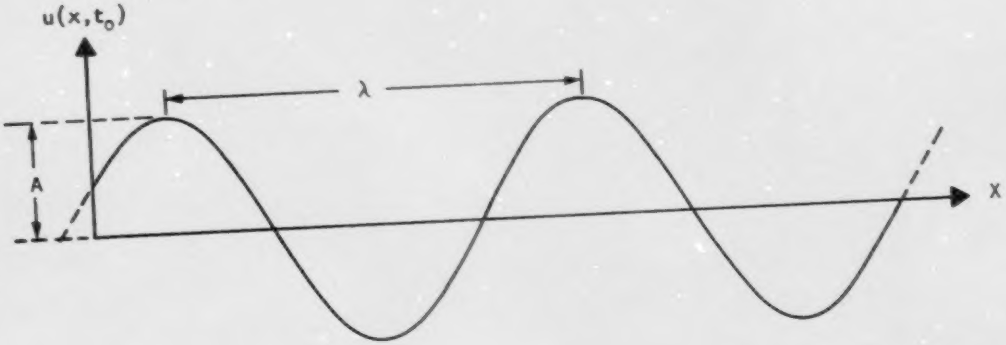


Figure 15-6b

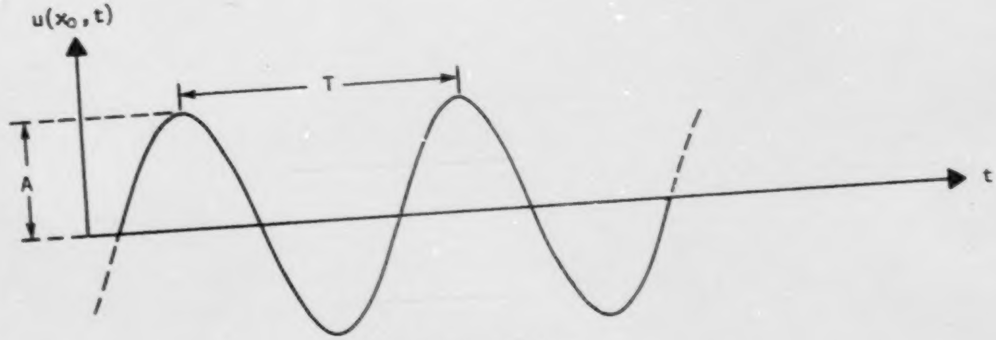


Figure 15-6c

In Figure 15-6b, we plot u of (5) versus x for fixed $t = t_0$, and in Figure 15-6c, we plot u of (5) versus t for fixed $x = x_0$. At a fixed moment of time, u is periodic in x; the wave form is repeated at intervals of length λ, which is why λ is called the wavelength or the space period.

Similarly at a fixed position on the x-axis, we observe that u is periodic in t : the wave form repeats itself at time intervals T called the *time period* or simply the *period*.

At a given time t_0 we obtain the wave form of Figure 15-6b. Later at $t_0 + \Delta t$ we obtain the same system of crests and valleys with each point of the wave shifted to $x + \Delta x$, where $\Delta x = v\,\Delta t$. Thus we may visualize the wave as traveling in the direction $\vec{i}$ of the x-axis.

The reciprocal of T is called the frequency of the source producing the wave and it is convenient to measure this frequency in units of 2π ; i.e., to use $\omega = \frac{2\pi}{T}$ where ω is called the *angular frequency*. Thus we rewrite (5) as

$$u = A\cos(kx - \omega t) \;. \tag{9}$$

Thus we identify the color parameter ω as the angular frequency of the wave associated with light of a single color.

The angular frequency ω is a property of the source of the waves and does not depend on the optical properties of the different media (characterized by different v) through which a wave passes; however, the wavelength $\lambda = \frac{2\pi v}{\omega}$ does depend on the medium. In general the phase velocity v is a function of ω, so that waves of different frequencies travel with different velocities $v(\omega)$ in the same material. Equivalently, since the index of refraction is defined as inversely proportional to v , we may rephrase the above in terms of $\mu(\omega)$. Taking the development until Equation (9) as applying to a medium with index of refraction $\mu = 1$ (*free-space* or vacuum), we replace kx for the more general case by

$$k\mu x = \frac{2\pi}{\lambda}\mu x = \frac{2\pi}{\lambda_\mu} x \tag{10}$$

where λ is the wavelength in the medium with $\mu = 1$ and λ_μ (a function of ω and the material) is the wavelength in the optical medium defined by $\mu(\omega)$.

The wavelength λ is the length factor which accounts for the differences in edge diffraction with color which we alluded to at the end of Section 15-2.

We could have introduced much of the above structure into the ray picture by associating the idea of phase (periodicity) with a ray. However, the wave picture is in general more fruitful for the usual visible phenomena. For convenience in the following, we may use a mixed terminology with the rays understood as the corresponding wave normals.

If light of a single color travels a distance L in vacuum, its phase has changed by kL. Corresponding to the unit sources of Section 15-4, the phase at a distance R along the ray from the source, differs from the phase at the source by kR. Similarly for the reflection problems of Section 15-2, the phase at P on the reflected ray in Figure 15-2d(i) differs by $k(L_1 + L_2)$ from the phase at S, and the phase at P on the ray in Figure 15-2g relative to the phase at $x = 0$ is given by $p = k(\xi + R) = k(\xi + \sqrt{(\xi - x)^2 - (\eta - y)^2})$. From the geometrical methods of constructing an eikonal (wave surface), we see that it is a curve (or surface) of constant phase (i.e., there is no phase difference between any two points on an eikonal), and we may label a particular eikonal by a particular value of p. Similarly for the elementary sources of Section 15-4, say the point source, we may surround the source (Figure 15-4a) with a set of eikonals, in this case spherical surfaces of particular radii R_n, corresponding to the particular phase differences kR_n.

Instead of working directly with $\cos p$ it is more convenient to manipulate formally with

$$e^{ip} = \cos p + i \sin p , \tag{11}$$

and take the real part $\mathrm{Re}(e^{ip}) = \cos p$ when we want to exhibit the periodic behavior explicitly.

Equation (11), and any operations of the calculus we may apply to it, can be justified in the following way.

The convergence of a power series can be defined in an obvious way for complex numbers $z = x + iy$. Then, if $\sum_{n=0}^{\infty} a_n x^n$ has radius of convergence R, the series $\sum_{n=0}^{\infty} a_n z^n$ will converge for those $z = x + iy$ with $(x^2 + y^2)^{1/2} = |z| < R$. It is then natural to extend the domain of the function $\exp : x \to e^x$ to the set of complex numbers by

$$\exp z : z \to e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} .$$

Then,

$$e^{ix} = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} = \sum_{n=0}^{\infty} \frac{i^n x^n}{n!}$$

$$= \sum_{n=0}^{\infty} \frac{i^{2n} x^{2n}}{(2n)!} + \sum_{n=0}^{\infty} \frac{i^{2n+1} x^{2n+1}}{(2n+1)!}$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} + i \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} ,$$

where we recognize

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \quad \text{and} \quad \cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} .$$

Thus we conclude that

$$e^{ix} = \cos x + i \sin x .$$

If we represent (11) in the complex plane we obtain the vector diagram (Argand diagram) of Figure 15-6d. As we progress along a ray (as we increase L), p increases and the tip of the vector of unit length describes a circle of unit radius. The projection of the tip on the x-axis (the real axis) is the oscillatory function $\cos p$; each time p increases by 2π (each time the tip describes a full circle), the x-projection goes through its maximum (+1) and minimum (-1) values. (The function e^{ip} is often called a phasor.) More generally, we work with

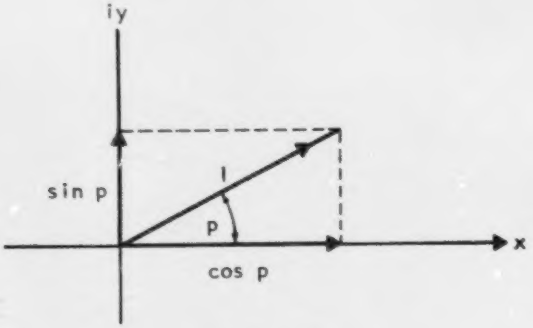


Figure 15-6d

(12) $$F(p) = Ae^{ip} , \qquad A > 0 ,$$

where A is the amplitude.

In subsequent applications we use the exponential form

(13) $$U(x,t) = Ae^{i(kx-\omega t)} ,$$

for which the function u of (9) is given by $\mathrm{Re}(U) = u$. We speak of (13) as a plane wave traveling in the direction of the x-axis.

In general, we consider waves which are not necessarily plane waves of the form (13), and A need not be constant. To tie in the present discussion with energy flux considerations of Section 15-4, we note that (in general) at distances from the source large compared to wavelength we may approximate A by a constant times $\sqrt{F}$, where F is the flux density introduced for the Kepler-Lambert principle. We write

(14) $$U \approx C\sqrt{F}\, e^{i(KL-\omega t)} ,$$

where L equals x or r, and where F in general depends on distance.

For the point source (or point scatterer) at the origin in three dimensions, we showed in Section 15-4 that $F = \frac{c}{r^2}$. We therefore write the corresponding wave as

(15) $$U = C_3 \frac{e^{i(kr-\omega t)}}{kr} , \quad r = \sqrt{x^2 + y^2 + z^2} ,$$

where we used kr (instead of r) in the denominator for convenience. Similarly for a line source (or line scatterer) along the z-axis, the wave corresponding to the flux density $F = \frac{c}{r}$ is

(16) $$U = C_2 \frac{e^{i(kr-\omega t)}}{\sqrt{kr}} , \quad r = \sqrt{x^2 + y^2} .$$

In the same sense that we interpret (13) as a wave traveling along the x-axis, we speak of (15) and (16) as waves traveling outward radially, or as outgoing waves. For the planar source (or planar scatterer) at $x = 0$, F is independent of distance, and the analog of (15) and (16) is

(17) $$U = \begin{cases} Ce^{i(kx-\omega t)} & \text{for } x > 0 \\ Ce^{-i(kx+\omega t)} & \text{for } x < 0 , \end{cases}$$

which we rewrite compactly as

(18) $$U = Ce^{i(k|x|-\omega t)} .$$

__Interference__. The concept of __interference__ was introduced into wave physics by Young. Later we shall discuss interference in detail but we mention it now to stress the most significant feature arising from associating a wave (or more specifically a phase) with light. The essentials are indicated in Figure 15-6e for scattering of a monochromatic plane wave by a screen containing two very narrow slits separated by $d \gg \lambda$. The waves from the two slits that arrive at (x,y) where $(x \gg d)$ have traveled different paths L_1 and L_2 , and therefore differ in phase by

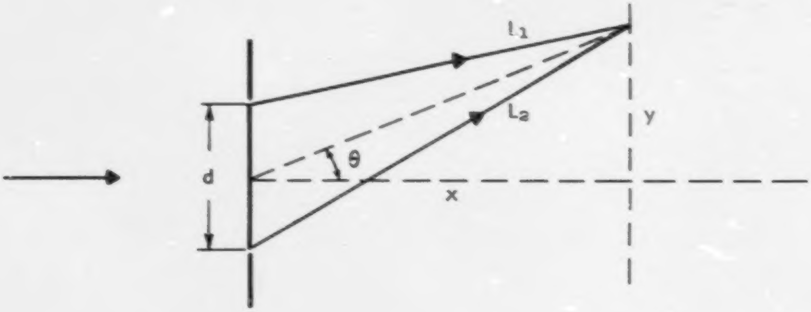


Figure 15-6e

(19)
$$\phi = k(L_2 - L_1) \approx kd \sin\theta = \frac{2\pi d}{\lambda} \sin\theta \ .$$

From (14), we may thus write the resultant wave at (x,y) in the form

(20)
$$U = U_1 + U_2 \approx W(1 + e^{i\phi}) \ ,$$

where we have used the fact that the fluxes from the two slits are equal to a first approximation and where we have absorbed $e^{ikL_1 - i\omega t}$ and other factors into W . (See Exercises 15-6, No. 1.) The corresponding energy flux or intensity is proportional to

(21)
$$F = |U|^2 = |W|^2 |1 + e^{i\phi}|^2 = |W|^2 2(1 + \cos\phi) \ .$$

For the moment we set $|W|^2$ equal to unity, and show the essentials of the interference effect in terms of the simpler function

(22)
$$F = |1 + e^{i\phi}|^2 = 2(1 + \cos\phi)$$

described vectorially in Figure 15-6f.

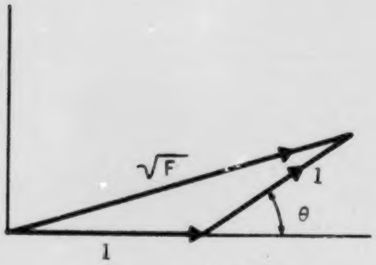


Figure 15-6f

We see that if $\phi = 0$ in (22) (i.e., $\theta = 0$ along the x-axis) then $F = 4$. (This corresponds essentially to a caustic of edge rays as discussed in Section 15-2; however, we now have much more structure for the description of light in the shadow region.) As we vary y keeping x fixed, the intensity F reaches the maximum value 4 when

(23) $$\phi = 2n\pi \quad , \quad n = 0 \ , \pm 1 \ , \pm 2 \ , \ \dots \ ,$$

and a minimum of zero when

(24) $$\phi = (2n + 1)\pi \ .$$

This behavior is clear from (22), and more graphically so from Figure 15-6f: if $\phi = 2n\pi$, then the two vectors of unit length point in the same direction along a straight line and their resultant is 2 ; if $\phi = (2n + 1)\pi$, then they point in opposite directions and cancel each other. The results for F with variation of ϕ are shown in Figure 15-6g.

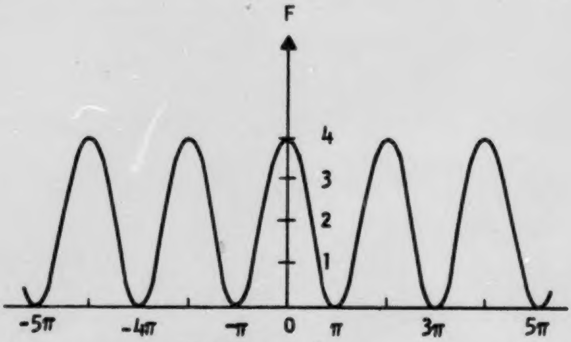


Figure 15-6g

Thus for a monochromatic wave (fixed λ) , a parallel screen on the shadow side of the slit-screen (the dashed line at x in Figure 15-6e) will show bright and dark bands: the bright bands or "fringes" corresponding to $\phi = 2n\pi$ are located on the screen at a distance x from the strip by

$$\frac{yd}{x} = 0 \; , \; \lambda \; , \; 2\lambda \; , \; \ldots = n\lambda \; ; \qquad y = \frac{n\lambda x}{d} \; . \tag{25}$$

i.e., when the path difference is an integral number of wavelengths. Similarly the dark fringes corresponding to $\phi = (2n + 1)\pi$ are located by

$$\frac{yd}{x} = \frac{\lambda}{2} \; , \; \frac{3\lambda}{2} \; , \; \ldots = (n + \tfrac{1}{2})\lambda \; , \qquad y = (n + \tfrac{1}{2})\lambda \, \frac{x}{d} \; . \tag{26}$$

We call (25) "constructive interference", and (26) "destructive interference."

If we use white light (a mixture of waves of different λ's), then on the axis at $y = 0$, we obtain a white central fringe; however, from (25), the side fringes are displaced from the axis in proportion to λ and we therefore see bands of different colors. (As we mentioned at the end of Section 15-2, analogous phenomena in the shadow region of a wide strip first noticed by Grimaldi could have led to the discovery of the periodic character of monochromatic light but did not.) Comparing (25) with experimental observations we find, from the displacement of the bands of light of different colors, that the wavelength for red light is about twice that of blue light, i.e.,

$$\lambda_r \approx 2\lambda_b \; , \tag{27}$$

and that the colors orange through yellow through green have wavelengths λ of length intermediate to that of red and blue.

In the remainder of the chapter we consider several elementary applications to scattering phenomena of the Huyghens-Newton-Young periodic wave theory of light. These applications are associated with Fraunhofer (1787-1826, an experimentalist), Fresnel (1788-1827, a theoretician), and Rayleigh (1842-1919, both).

Fraunhofer Diffraction by a Slit. We next apply the wave model to Fraunhofer diffraction of a plane wave by a slit of width $2a$ in a perfectly reflecting plane as in Figure 15-6h. We take the origin at the center of the slit.

We write the incident wave as

$$U_i(x,y,t) = U_i = e^{i(kx-\omega t)} \tag{28}$$

and interpret (28) as a wave of unit flux density traveling in the x direction. Using [Hu] implicitly, we regard U_i as exciting wavelets in the plane of the aperture, and specify a wavelet originating at $x = 0$, $y = 0$ by the

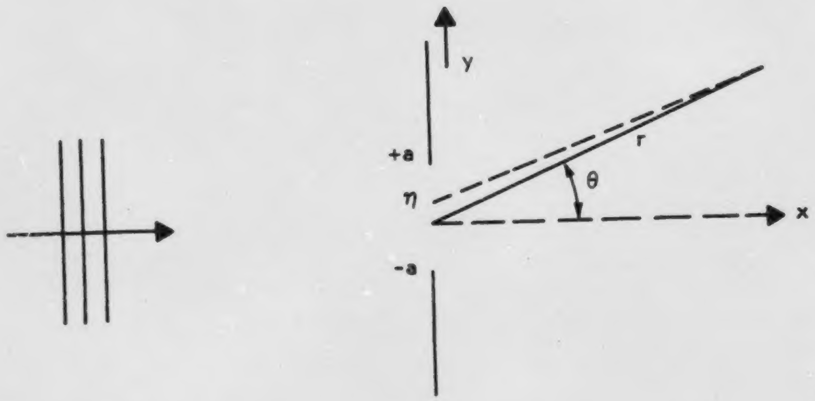


Figure 15-6h

elementary outgoing wave form $E\frac{e^{i(kr-\omega t)}}{\sqrt{kr}}$ as in (16). Similarly for a wavelet originating at $(0,\eta)$ as in Figure 15-6h we use

$$u = C\,\frac{e^{i(kR-i\omega t)}}{\sqrt{kr}}\ ,\ R = \sqrt{r^2 - 2r\eta\sin\theta + \eta^2}\ . \tag{29}$$

Every point of the line $y = \eta$, $x = 0$, for $-a \le \eta \le a$ (every line element of the slit) corresponds to a wavelet of the form (29). We represent the net effect of all such wavelets at a distant point $\vec{r}$ by the integral

$$U = \int_{-a}^{a} u(\eta)d\eta\ . \tag{30}$$

Restricting consideration to $r \gg a$, we approximate R in the exponent by

$$R \simeq r - \eta\sin\theta\ . \tag{31}$$

In the denominator we use simply $R \approx r$, because $|U|$ is much less sensitive to changes in the denominator than to changes of the phase. (From Figure 15-6f we see that a slight change of the magnitudes of the two vectors has little effect compared to a comparable change in the phase difference ϕ .) Thus (30) reduces to

$$U \approx C\,\frac{e^{i(kr-\omega t)}}{\sqrt{kr}}\,g(\theta) \tag{32}$$

where

(33) $$G(\theta) = \int_{-a}^{a} e^{ik\eta \sin\theta} d\eta ,$$

i.e., U is an elementary cylindrical wave (source at the origin) as in (25), times a function of angles **$G(\theta)$** (the scattering amplitude). Thus for $\theta = 0$ or π, we have

(34) $$G(0) = G(\pi) = \int_{-a}^{a} d\eta = 2a ,$$

where $2a$ is the width of the strip. For the other angles, we integrate the exponential and obtain

(35) $$G(\theta) = \frac{e^{ika \sin\theta} - e^{-ika \sin\theta}}{ik \sin\theta} = \frac{2i \sin(ka \sin\theta)}{ik \sin\theta}$$
$$= 2a\left[\frac{\sin(ka \sin\theta)}{ka \sin\theta}\right] \equiv S\,\Gamma(\theta) ,$$

where S is the width of the strip, and Γ is an oscillatory function with zeros at $ka \sin\theta = n\pi$. (See Exercises 15-6, No. 2.)

Λ(ii) Rayleigh-Born scattering by a sphere.

As another illustration we consider Rayleigh-Born scattering by a sphere of radius a whose optical properties differ only very slightly from the free space in which it is imbedded (a "tenuous" scatterer). We use the geometry of Figure 15-6i with center of the sphere at the origin, and take the plane wave

(36) $$U_i = e^{i(kz-\omega t)}$$

as the incident field.

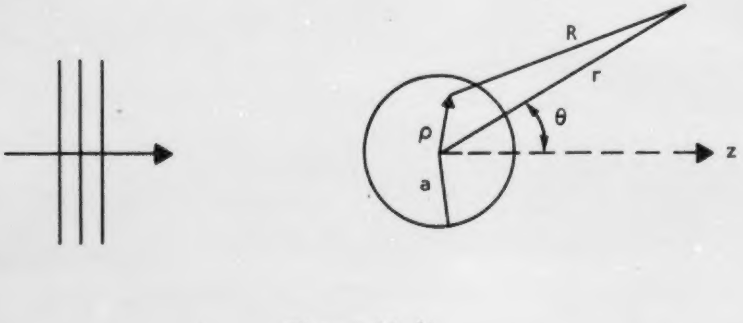


Figure 15-6i

We regard the sphere as made up of elementary sources of spherical wavelets, such that the source at the origin produces a wavelet $C \frac{e^{i(kr-\omega t)}}{kr}$ as in (15). The elementary source at the position $\vec{\rho} = (\xi,\eta,\zeta)$ excited by the incident field $U_i = e^{i(k\zeta-\omega t)}$ produces an effect at $\vec{r} = (x,y,z)$ described by

(37) $$U = \frac{C}{kR} e^{ik(\zeta+R)-i\omega t} ,$$
$$R = \sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2} = \sqrt{(r^2 + \rho^2 - 2\vec{r}\cdot\vec{p})} ,$$

where the phase is chosen to agree with the phase of the incident wave at $\vec{\rho}$ $(R = 0)$. The net effect is represented by the volume integral of u over the sphere of radius a :

(38) $$U = \int u(\rho)dV(\vec{\rho}) .$$

(We define the volume integral as the limit of sums $\sum u(\vec{\rho}_n)\Delta V_n$ where the volume is subdivided into elements with volumes ΔV_n, where $\vec{\rho}_n$ lies in the corresponding volume element, and, the limit is taken as the maximum diameter of the volume elements approaches 0.)

Essentially as for the slit problem we restrict consideration to $r \gg a$, and approximate R in the exponent of u by

(39) $$R \approx r - \frac{\vec{r}}{r}\cdot\vec{\rho} ;$$

in the denominator, we use simply $R \approx r$. Thus (38) becomes

(40) $$U \approx C \frac{e^{i(kr-\omega t)}}{kr} G(\theta) ,$$

(41) $$G(\theta) = \int e^{ik(\zeta-\vec{\rho}\cdot\vec{r}/r)}dV ,$$

i.e., U is the product of an elementary spherical wave outgoing from the origin as in (15) and a scattering amplitude $G(\theta)$ which is independent of r but depends on the angle of observation θ. For present purposes we do not evaluate the integral in (41) but merely show how it is analogous to (35).

In the forward direction $\theta = 0$, we have $\vec{r} = (0,0,z)$ so that $\vec{\rho}\cdot\frac{\vec{r}}{r} = \zeta$ and (41) reduces to

(42) $$G(0) = \int dV = V ,$$

where V is the volume of the sphere. For other directions, we write

(43) $$G(\theta) = VJ(\theta)$$

where $J(\theta)$ is an oscillatory function analogous to $\Gamma(\theta)$ of (35). It can be shown that

$$J(\theta) = \frac{\sin X}{X^3} - \frac{\cos X}{X^2}\ , \quad X = 2ka \sin\frac{\theta}{2}\ ,$$

which approaches $\frac{1}{3}$ if $\theta = 0$, and also approaches $\frac{1}{3}$ if $ka = \frac{2\pi a}{\lambda}$ approaches 0. We may use the fact that U is proportional to $\frac{V}{r}$ for a variety of different scatterers to further the discussion on the color of the sky that we touched on in Chapter 9. We consider this in the following, which supplements the present section and Section 4.

∧(iii) Rayleigh Scattering.

In 1871, Rayleigh developed a mathematical model to account for the blue color of the sky based essentially on a scattering integral similar to (38). He obtained the constant c explicitly and showed that for small $\frac{a}{\lambda}$ (small scatterers) U was proportional to $\frac{V}{\lambda^2} r$. The ratio of the scattered flux density at $\hat{r}$ to that incident on the scatterer (see Figure 15-6j) is proportional

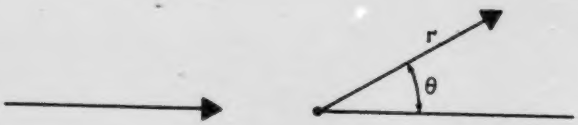


Figure 15-6j

to $|U|^2$ and can thus be written

(44) $$\frac{F}{F_0} = \frac{g^2 V^2}{r^2 \lambda^4}\ ,$$

where g^2 does not depend on length factors. Rayleigh also obtained the same dependence on λ from dimensional analysis by starting with

(45) $$\frac{F}{F_0} = \frac{BV^2}{r^2}$$

(as obtained in our primitive discussion of (38)) and requiring that B not depend on any length dimension of the scatterer: since F and F_0 are different values of the same physical quantity, their ratio is free of units or dimensions; since $\frac{V^2}{r^2}$ has the dimension of length to the fourth power, and since λ is the only available length, it then follows that B is proportional to $G\lambda^{-4}$ and (45) reduces to the form (44).

From Equation (3) of Section 9-3 (i.e., from $\lambda_r = 2\lambda_b$, where λ_r and λ_b are the wavelengths of red and blue components of sunlight), and from (44) it follows that

$$\frac{F(\lambda_b)}{F(\lambda_r)} = \frac{\lambda_r{}^4}{\lambda_b{}^4} = \frac{(2\lambda_b)^4}{\lambda_b{}^4} = 16 \ . \tag{46}$$

Thus, if a beam of white light is incident on a small scatterer as in Figure 15-6j, the blue component of white light is scattered sixteen times as strongly as the red; i.e., the flux ratio of the blue and red components observed at an angle θ from the direction of incidence is given by (46).

Equation (44) specifies the scattered flux density at a point $r(\theta)$ as in Figure 15-6j. The total flux scattered through a complete spherical surface around the scatterer (as discussed in Section 4) is thus independent of r. If $F_0 = 1$, the total scattered flux (the scattering cross section) may be written as

$$Q = \frac{A}{\lambda^4} \tag{47}$$

in order to emphasize the dependence on λ.

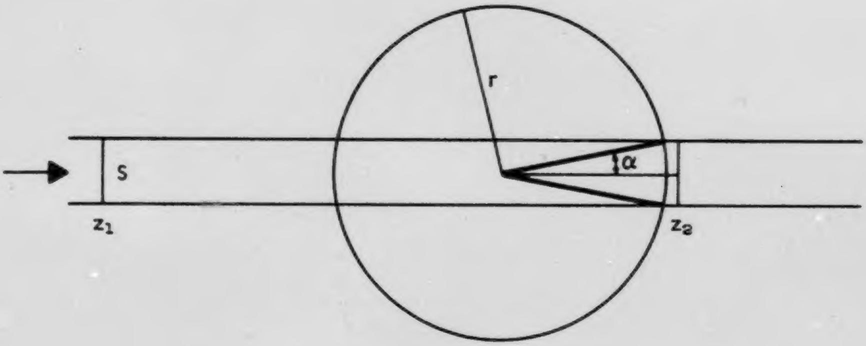


Figure 15-6k

Let us now visualize an incident beam of rays flowing through a tube of cross-sectional area S as in Figure 15-6k, and apply the Kepler-Lambert flux principle [KL]. The flux through S at z_1 is I_1S, and from [KL] this equals the flux I_2S through the surface S at z_2 plus the flux through the incomplete spherical surface consisting of the sphere minus the cap α. If we neglect the "hole" in the sphere, we approximate the scattering out of the beam by Q, and obtain

$$I_1S = I_1Q + I_2S \;. \tag{48}$$

Since Q is independent of r, the difference between the initial and final values of the flux along a parallel beam intercepted by a scatterer anywhere along the beam is

$$I_2S - I_1S = -I_1Q \;. \tag{49}$$

If there are N such scatterers in the beam, then under appropriate restrictions we may use (49) with the right-hand side multiplied by N to approximate the net effect:

$$(I_2 - I_1)S = -NI_1Q \;. \tag{50}$$

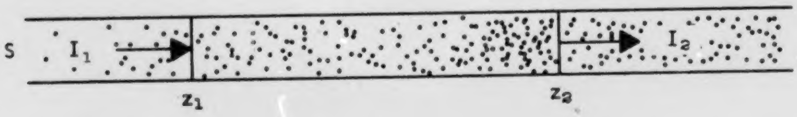


Figure 15-6ℓ

If there are n scatterers in unit volume in the geometry of Figure 15-6ℓ, then we have $N = nS(z_2 - z_1)$, and (50) reduces to

$$I_2 - I_1 = -nQI_1(z_2 - z_1) \;. \tag{51}$$

In the limit as z_1 approaches z_2, we obtain

$$\frac{dI}{dz} = -nQI \;, \tag{52}$$

and consequently

$$I = I_0e^{-nQz} \tag{53}$$

with I_0 as the value at say $z = 0$.

From (47), we have $Q = \dfrac{A}{\lambda^4}$, so that

$$I = I_0 e^{-n(A/\lambda^4)z} , \tag{54}$$

which is the form considered in Chapter 9. Thus, as (45) accounts for the blue of the sky in directions other than towards the sun, (54) accounts for the redness of the clouds illuminated by sunlight near dusk. (See the discussion of (6) in Section 9-3.)

Equation (45) gives the scattered intensity for one scatterer excited by a wave of unit intensity. If one scatterer of a collection is at a distance z from the entrance face of the region of scatterers (as in Figure 15-6m),

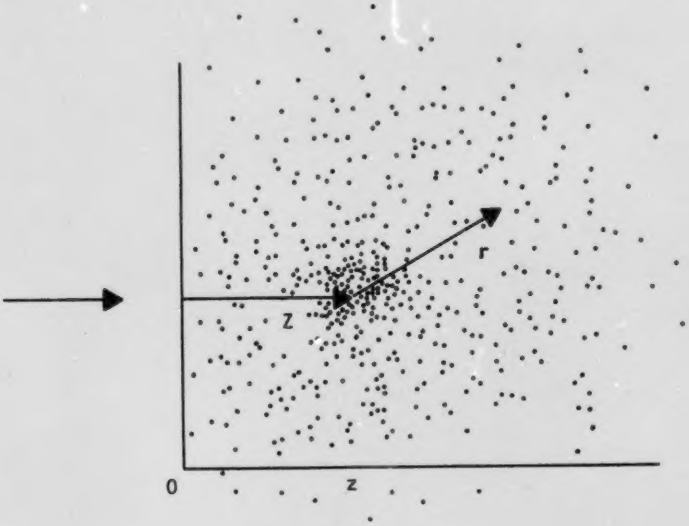


Figure 15-6m

then we multiply (45) by $I(z) = I_0 e^{-nQz}$ of (53) to account for the intensity loss of the excitation that reaches it. Similarly, if we observe a scattered ray from this scatterer at a distance r from its center, we incorporate an additional factor e^{-nQr} to account for the additional loss. Thus, the scattered intensity for one scatterer as in Figure 15-6m, becomes

$$F = \frac{g^2 V^2}{r^2 \lambda^4} I_0 e^{-n(A/\lambda^4)(z+r)} \tag{55}$$

where $z + r$ is the total ray path within the region of scatterers.

Let us rewrite (22) for $z + r = 1$ as

(56)
$$F = \frac{B}{\lambda^4} e^{-D/\lambda^4}$$

a form that shows that F vanishes for both $\lambda \sim 0$ and $\lambda \sim \infty$, and has a maximum at a definite value of λ, say $\lambda = \Lambda$. Differentiating (56) with respect to λ, we obtain

(57)
$$\frac{dF}{d\lambda} = \frac{dF}{d\lambda^{-4}} \frac{d\lambda^{-4}}{d\lambda} = Be^{-D\lambda^{-4}}(1 - \lambda^{-4}D)(-\frac{4}{\lambda^5})$$

which vanishes for the wavelength

(58)
$$\lambda^4 = \frac{1}{D} = \Lambda^4$$

corresponding to a maximum scattered intensity

(59)
$$F_\Lambda = \frac{B}{\Lambda^4} e^{-1} = \frac{B}{\Lambda^4 e} .$$

Dividing (56) by (59) we write the scattered intensity as

(60)
$$F = F_\Lambda \frac{\Lambda^4}{\lambda^4} e^{1-\Lambda^4/\lambda^4} ,$$

so that F is expressed in terms of the maximum value F_Λ and the corresponding wavelength Λ. This simple model applied to skylight gives a maximum wavelength in the blue-green region of the sun's spectrum.

Exercises 15-6

1. (a) Show that (19) is valid to first order in the parameter $\epsilon = \frac{d}{x}$.

   (b) Show that the error in (20) is at most first order in ϵ.

2. Determine the extrema of the scattering amplitude $G(\theta)$ for a slit, Equation (33), by calculating the extrema of $f : x \to \frac{\sin x}{x}$. (The graph is given in Figure A 10, p. 665).

15-7. Method of Stationary Phase.

We are now in a position to bridge the gap between the geometrical optics ray procedures of the early sections and the elementary wave procedures of Section 15-6. We do so initially by considering diffraction by a strip. We work with the wave forms of Equations (28), (29), and (30) of Section 15-6 and the geometry of Figure 15-7a. Thus we take the incident wave proportional to

$$U_i = e^{ikx} , \tag{1}$$

the wavelet from a secondary line source on the strip as

$$u = \frac{ce^{ikR}}{\sqrt{kR}} , \quad R = \sqrt{x^2 + (y - \eta)^2} , \tag{2}$$

and represent the net effect of the wavelets at r as the integral

$$U = \int_{-a}^{a} u(\eta)d\eta = c\int_{-a}^{a} \frac{e^{ikR}}{\sqrt{kR}}\, d\eta . \tag{3}$$

In each of these wave forms we have suppressed the time factor $e^{-i\omega t}$ for brevity. To obtain the actual wave forms we must multiply above by $e^{-i\omega t}$ and then take the real part of the result.

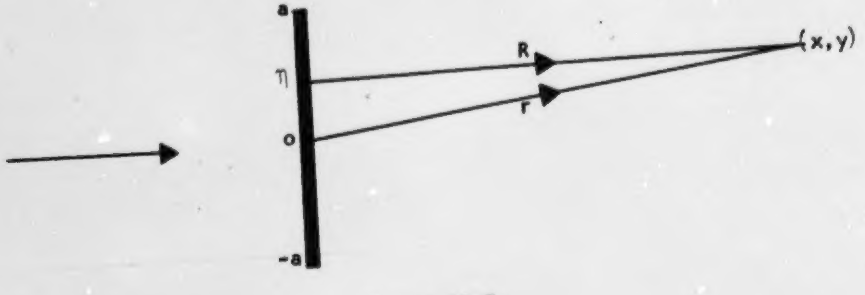


Figure 15-7a

The situation in Figure 15-7a is analogous to that of Figures 15-6e, f, h, and if we assume $R \gg a = |\eta|_{max}$ we obtain the same forms as (32) to (35) of Section 15-6. Thus if we expand $\frac{R}{r}$ to first order in $\frac{\eta}{r}$ (as before), we obtain $R \approx r - \eta \sin\theta$, and consequently the previous procedure yields

(4)
$$\begin{aligned} U &\approx \frac{ce^{ikr}}{\sqrt{kR}} \int_{-a}^{a} e^{ik\eta \sin\theta}\, d\eta \\ &= \frac{ce^{ikr}}{\sqrt{kR}} \left[\frac{e^{ika\sin\theta} - e^{-ika\sin\theta}}{ik\sin\theta} \right] \\ &= \frac{ce^{ikr}}{\sqrt{kR}}\, 2a \left[\frac{\sin(ka\sin\theta)}{ka\sin\theta} \right] . \end{aligned}$$

Thus except that the constant c may differ in the two cases, the present result (the Fraunhofer approximation for scattering or diffraction by a strip) is the same as that for the slit. (The relation of the result for the strip to that of the slit shown by the above is a special case of what is called Babinet's principle.)

We introduce $\gamma = ka \sin\theta$ to represent the phase difference between the rays from the center and from an edge to the observation point, and write $\Gamma = \frac{\sin\gamma}{\gamma}$ (the "Fraunhofer pattern factor"). The principal maximum of Γ corresponds to $\gamma = 0$ (i.e., to the forward and back directions, $\theta = 0$ and π). The secondary maxima of Γ occur at the zeros of $\tan\gamma - \gamma$, which are given approximately by $\gamma \approx 1.43\pi$, 2.46π, 3.47π, and, for larger values, by $\gamma \approx (m + \frac{1}{2})\pi$. For the first three of these zeros of $\frac{d\Gamma}{d\gamma}$, we have $\Gamma \approx -0.22$, 0.13, -0.09 respectively. The zeros of Γ correspond to $\gamma = n\pi$. The angular half-width of the principal maximum (obtained from the position of the first zero, $\gamma = ka\sin\theta = \pi$) is $\sin\theta \approx \theta = \frac{\pi}{ka} = \frac{\lambda}{2a}$.

The form (4) is restricted to $r \gg a$. In order to consider situations where r and a are comparable in magnitude, or $r < a$, we use a different approximation for R in (2). Thus we now relax the requirement $r \gg a$ and assume instead that we restrict the direction of observation to the neighborhoods of the forward scattered (the direction of incidence) and back scattered directions. More explicitly, we assume $(y - \eta)^2 \ll x^2$, and use

(5)
$$R = \sqrt{x^2 + (y - \eta)^2} \approx x + \frac{(y - \eta)^2}{2x} .$$

We use (5) in the exponent of (3), but in the denominator (essentially as for Equation (32) of Section 15-6) we use only the leading term $R \approx x$. Thus

(6)
$$U \approx \frac{ce^{ikx}}{\sqrt{kx}} \int_{-a}^{a} e^{ik(y-\eta)^2/2x} d\eta .$$

The integral (6) describes the so-called Fresnel diffraction by a strip.

After we have analyzed the behavior of (6), we shall also treat the corresponding problem for scattering by a circular cylinder. A limiting case of the result we shall obtain will correspond to our geometrical optics results of Sections 15-2 and 4 in the form

$$U = \frac{ce^{ikL_H}}{\sqrt{RL_H''}} = C\sqrt{F}\,e^{ikL_H} , \tag{7}$$

where L_H is the ray path obtained previously by using Hero's principle [H'] , L_H'' is its second derivative with respect to a parameter along the circle, and F is the flux density obtained by the Kepler-Lambert principle [KL] .

Our derivation of (7) from (3) will obviate the special assumptions of geometrical optics. In particular, we shall not have to assume Hero's "principle of the extremum path." Statements such as "nature prefers an extremum", "nature abhors a vacuum", etc., may be useful aids to memory, but "nature" neither "prefers" nor "abhors", and such statements make our mysticism too explicit. We will show that Hero's principle [H'] is merely a consequence of an approximate evaluation of an integral.

The approximation procedure is known as the *method of stationary phase*. It was introduced by Kelvin as a mathematical *method* for approximating a class of integrals that come up frequently in wave problems. Here we are concerned with integration over an arc, $U = \int uds$, essentially as in (3).

Intuitively, the method is based on Young's concept of interference, and the essentials were discussed for Figure 15-6f. A complex number $Ae^{i\phi}$ may be represented as a vector of length A and direction angle (phase) ϕ on an Argand diagram and the resultant of a set of numbers $A_n e^{i\phi_n}$ is simply the vector sum $Te^{i\tau} = \Sigma A_n e^{i\phi_n}$ depicted in Figure 15-7b or c. If the phase angles are all the same, then the elementary vectors all point along a straight line and $T = \Sigma A_n$: we say that the vectors reinforce each other, or that the elementary waves "interfere constructively." On the other hand, if the angles are such that the nose of the last vector ends up at the tail of the first to form a closed polygon, then $T = 0$; we say that the vectors cancel, or that the elementary waves "interfere destructively." In the situation shown in Figure 15-7b, the phase ϕ_n changes slightly (varies slowly) with increasing n and the resultant magnitude T is accordingly large. In the situation shown in Figure 15-7c, the variation of ϕ_n with n is large and T is small. The quantities A_n and ϕ_n may depend on a parameter θ , and the magnitude $T(\theta)$ of the resultant given by

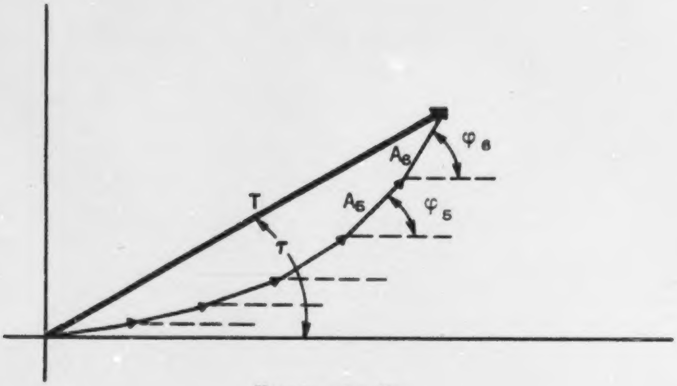


Figure 15-7b

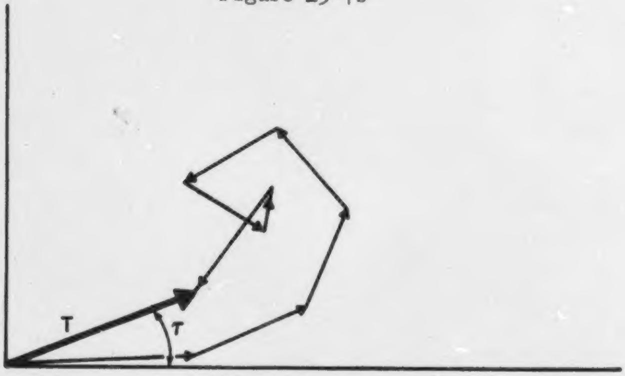


Figure 15-7c

(8) $$T(\theta)e^{i\tau(\theta)} = \sum_{n=1}^{N} A_n(\theta)e^{i\phi_n(\theta)}$$

may assume any value between 0 and ΣA_n with variation of θ . Similarly we may use the same idea for an integral of the form

(9) $$\int_{\eta_1}^{\eta_2} A(\eta,\theta)e^{i\phi(\eta,\theta)}d\eta \ .$$

The integral of (4) is of the above form, and the series of maxima and minima shown by the resultant can and will be interpreted graphically by means of a vector diagram such as Figure 15-7b. Similarly for the integral in (6). (See Exercises 15-7, No. 1.)

In (9), if for a fixed θ , $\frac{d\phi}{d\eta} = 0$, when $\eta = \eta_s(\theta)$ then we say the phase is stationary at η_s ; in the vicinity of η_s , the phase ϕ changes slowly with η ; the situation is analagous to that of Figure 15-7b, (not Figure 5-7c), and we expect the contribution to the resultant $T(\theta)$ to be large. To prepare for the mathematical discussion of the general method, we first consider the strip problem without using the method explicitly. The results we obtain initially be relatively familiar procedures will provide a basis for introducing the concepts needed for our subsequent more general discussion.

Fresnel Diffraction. We may express (6) in terms of the tabulated Fresnel integral

$$\mathcal{F}(\eta_1) = \int_0^{\eta_1} e^{i\pi\eta^2/2}\, d\eta = -\mathcal{F}(-\eta_1) \ , \tag{10}$$

whose path in the complex plane (i.e., the trace of the point $\mathcal{F}(\eta_1)$ as a function of η_1) generates Cornu's spiral (Exercises 11-5, No. 6(c)) of Figure 15-7d. The magnitude $|\mathcal{F}(\eta_1)| = T(\eta_1)$ is a damped oscillatory function of η_1. Figure 15-7d for (10) is a continuous analog of such cases of

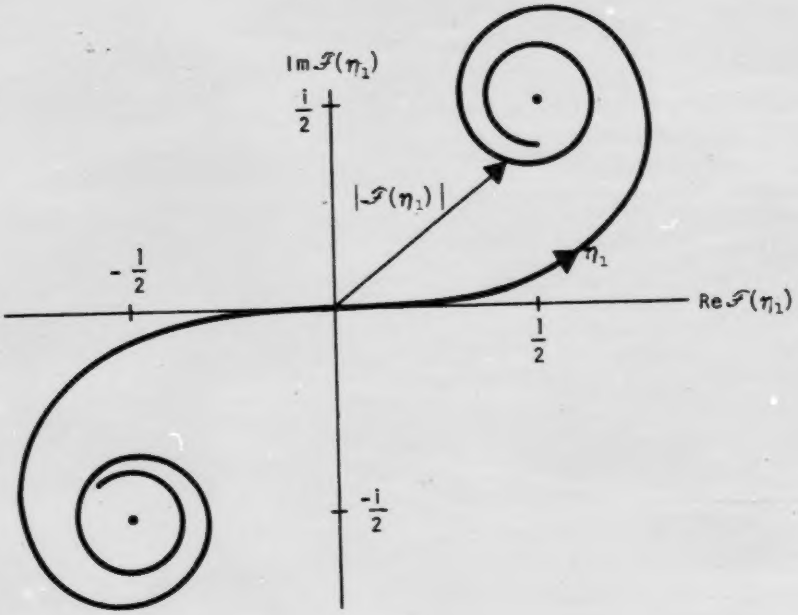


Figure 15-7d

discrete vectors shown in Figure 15-7b and Figure 15-7c. The phase changes slowly in the vicinity of $\eta_1 = 0$ and then changes with increasing rapidity, so that the contribution to the resultant (η_1) is greatest near $\eta_1 = 0$. Although the curve of Figure 15-7d has infinite length, a η_1 increases the curve spirals inward to a limit at the point $\mathcal{F}\,(\omega) = \frac{1}{2} + \frac{i}{2}$; the resultant $T = |\mathcal{F}|$ approaches the limit $\frac{1}{\sqrt{2}}$ as $\eta_1 \sim \infty$.

We now consider $\mathcal{F}\,(\eta)$ analytically. For large values of η_1, we use

$$(11) \qquad \mathcal{F}\,(\eta_1) = \left[\int_0^\infty - \int_{\eta_1}^\infty\right] e^{i\pi\eta^2/2}\, d\eta = \mathcal{F}\,(\infty) - \frac{1}{\pi i}\int_{\eta_1}^\infty \frac{1}{\eta}\, d(e^{i\pi\eta^2/2}) \; .$$

Integrating the second term by parts, we develop (11) as the series

$$(12) \qquad \mathcal{F}\,(\eta_1) = \mathcal{F}(\infty) + \frac{e^{i\pi\eta_1{}^2/2}}{i\pi\eta_1}[1 + \frac{1}{i\pi\eta_1{}^2} + \ldots] \; .$$

Thus to lowest order the error in using the leading term $\mathcal{F}(\infty)$ may be bounded in proportion to $\frac{1}{\eta_1}$. For the leading term itself, we accept without proof

$$(13) \quad \mathcal{F}(\infty) = \int_0^\infty e^{i\pi\eta^2/2}\, d\eta = \frac{1 + i}{2} = \frac{1}{2}[\cos\frac{\pi}{4} + i \sin\frac{\pi}{4}] = \frac{e^{i\pi/4}}{\sqrt{2}} \; .$$

(The proof lies outside the range of the text.)

On the other hand for small values of η, we expand the integral of (10) as a power series in η and integrate term by term:

$$(14) \qquad \mathcal{F}\,(\eta_1) = \int_0^{\eta_1} [1 + \frac{i\pi\eta^2}{2} + \ldots]d\eta = \eta_1 + \frac{i\pi\eta_1{}^3}{6} + \ldots \; .$$

In terms of (10) we rewrite the integral (6) for Fresnel diffraction by a strip as

$$(15) \qquad U = \sqrt{\pi}\,\frac{ce^{ikx}}{k}\, I \; ,$$

$$(16) \qquad I = \int_{\eta_-}^{\eta_+} e^{i\pi\eta^2/2}\, d\eta = \mathcal{F}(\eta_+) - \mathcal{F}\,(\eta_-) \; ,$$

$$\eta_\pm = \sqrt{\frac{k}{\pi x}}\,(\pm\, a - y) \; .$$

Before applying this result to the strip problem, let us first apply it to an "infinitely wide slit", and determine c for the Huyghens' free-space wavelets that simply serve to regenerate the incident wave. For the limiting case $ka \sim \infty$, we have $\eta_{\pm} \sim \pm\infty$. Since $\mathcal{F}$ is an odd function (10), from the limit of (13) we obtain

$$I \sim \mathcal{F}(\infty) - \mathcal{F}(-\infty) = 2\mathcal{F}(\infty) = \sqrt{2}\, e^{i\pi/4} , \tag{17}$$

and consequently

$$U \sim c\, \frac{\sqrt{2\pi}}{k}\, e^{i\pi/4}\, e^{ikx} . \tag{18}$$

But an "infinitely wide slit" means no obstruction, so that U of (18) must equal the incident wave e^{ikx}. Consequently, for the elementary Huyghens' sources is

$$c = \frac{k}{\sqrt{2\pi}}\, e^{-i\pi/4} . \tag{19}$$

More generally, if we are dealing with secondary sources on a scatterer excited by the incident wave, we may write

$$c = \frac{k}{\sqrt{2\pi}}\, e^{-i\pi/4}\, g , \tag{20}$$

where g may depend on the material of the scatterer, and on directions. Thus we may rewrite (15) as

$$U = g\, \frac{e^{-i\pi/4}}{\sqrt{2}} [\mathcal{F}(\eta_+) - \mathcal{F}(\eta_-)] e^{ikx} , \quad \eta_{\pm} = \sqrt{\frac{k}{\pi x}}\, (\pm a - y) . \tag{21}$$

We now apply (21) to scattering by a strip as in Figure 15-7a. There are essentially three different ranges of y that we consider.

In Figure 15-7e we specify three different ranges of y at a fixed value of x, say three different portions of a screen placed parallel to the strip. We will use (21), (12), and (14) to obtain explicit approximations of U for the three ranges of y corresponding to the braces shown in the figure. The range of y_2 is centered on the geometrical projection of the edge of the strip (equivalently the neighborhood of the shadow boundary $y = a$), and the range of y_1 includes much of the geometrical projection (the shadow) of the strip on the screen. The range of y_1 corresponds to $\eta_+ \gg 1$ and $-\eta_- \gg 1$, and includes $y = 0$ as the special case $\eta_+ = -\eta_- \gg 1$; the range of y_2 corresponds to $\eta_+ \approx 0$ and $-\eta_- \gg 1$; the range of y_3 corresponds to $-\eta_+ \gg 1$ and $-\eta_- \gg 1$. If we replace x, y by $|x|$, $|y|$, then the results will apply not only for the three sets of points (x, y_i) in the first quadrant shown in Figure 15-7e, but also to the sets obtained in the other

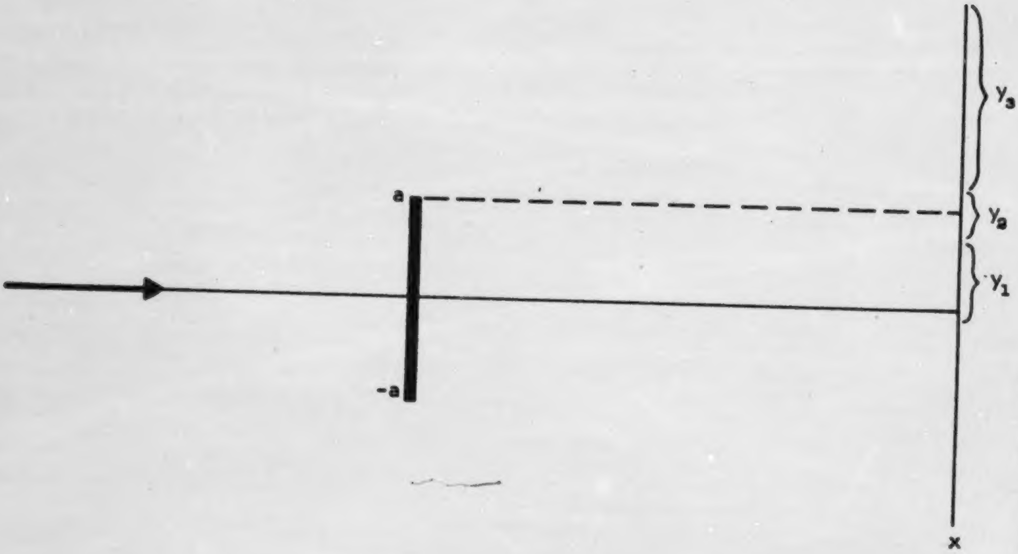


Figure 15-7e

three quadrants by reflecting the three given sets in the x-axis and in the y-axis. If, say, $\eta_+ = \eta_1 \gg 1$, we approximate the integral by the first two terms of (12) obtained by using (13),

$$\mathscr{F}(\eta_1) \sim \frac{e^{i\pi/4}}{\sqrt{2}} + \frac{e^{i\pi\eta_1^{\,2}/2}}{i\pi\eta_1}\,, \qquad \eta_1 \gg 1 \tag{22}$$

however, if η_1 is very small then we use only the leading term of (14).

$$\mathscr{F}(\eta_1) \approx \eta_1\,, \qquad \eta_1 \approx 0\,. \tag{23}$$

At a point (x,y_1) or more generally at the four points $(|x|,|y_1|)$, we have $\pm\,\eta_\pm = \sqrt{\frac{k}{x|x|}}\,(a \pm y) \gg 1$; substituting (22) into (21), we then obtain

$$\begin{aligned} U &= g\,\frac{e^{ik|x|}}{\sqrt{2}}\,e^{-i\pi/4}[\mathscr{F}(\eta_+) + \mathscr{F}(-\eta_-)] \\ &\approx g e^{ik|x|}\left\{1 - \sqrt{\frac{|x|}{2\pi k}}\,e^{+i\pi/4}\left[\frac{e^{ik(a-y)^2/2|x|}}{a-y} + \frac{e^{ik(a+y)^2/2|x|}}{a+y}\right]\right\} \end{aligned} \tag{24}$$

where g is not necessarily the same for the forward scattered direction $|x| = x$ as the back scattered direction $|x| = -x$. Essentially as for (5), we require $\left|\frac{(a \pm y)}{x}\right| \ll 1$; subject to this we see that in the regime limit of small $\left|\frac{x}{k(a \pm y)^2}\right|$

(25) $$U \approx g e^{ik|x|} .$$

Thus in this limit the strip is a one-dimensional secondary source. Taking into account that U is in general a function of direction, we write

(26) $$U_{\pm} \sim g_{\pm} e^{\pm ikx} \qquad \text{for} \qquad \operatorname{sgn} x = \pm 1 .$$

In the forward direction, we require $g_+ = -1$ in order for a geometrical shadow to exist in the sense of Section 15-2, i.e., $U_+ \sim -e^{+ikx}$. Similarly for the case of a perfectly reflecting strip, we require $|g_-| = 1$ in order that the ratio of the reflected to incident flux density $\left|\frac{\bar{U}_-}{U_i}\right|^2$ equals unity as before. Thus we have

(27) $$\begin{aligned} U_+ &\sim -e^{+ikx} , && \text{for } x > 0 , \\ U_- &\sim g_- e^{-ikx} = e^{i\delta} e^{-ikx} , && \text{for } x < 0 , \end{aligned}$$

where δ is a real number determined by the material of the strip (or, described mathematically, by the boundary conditions). For present purposes we take $\delta = 0$, so that

(28) $$U_- \sim e^{-ikx} \quad \text{for} \quad x < 0 ;$$

physically this corresponds for example to a waterwave or a sound wave incident on a rigid immovable strip, and also (subject to additional conditions) for the reflection of an electromagnetic wave from a metal strip.

We introduce the abbreviation

(29) $$u_e(Y) = -\sqrt{\frac{|x|}{2\pi k}}\, e^{i\pi/4} \frac{e^{ik[|x|+Y^2/2|x|]}}{Y} ,$$

so that we may rewrite (24) as

(30) $$U_{\pm} \approx \mp [e^{\pm ikx} + u_e(a - y) + u_e(a + y)] .$$

The present results apply for the geometry of Figure 15-7f. From (5) we see that $R_{\pm} \approx x + \frac{(a \mp y)^2}{2x}$ so that the exponents of $u_e(a \mp y)$ are approximations of $kR_{\pm}$. The factors of $kR_{\pm}$ are the phase changes of waves traveling

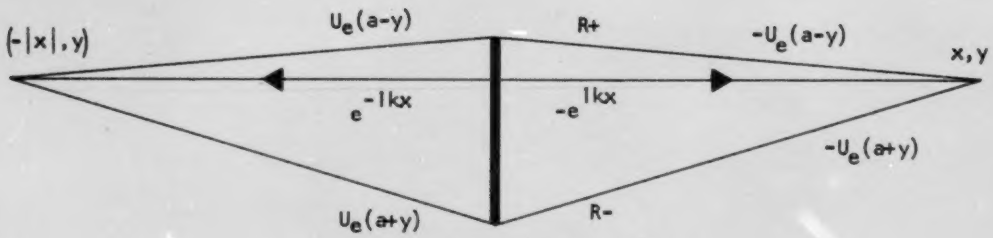


Figure 15-7f

from the edges of the strip $(0, \pm a)$ to the observation point (x,y) so that we may interpret $u_e(a \pm y)$ as the edge waves corresponding to the edge rays we introduced in Section 15-2. Thus the normals of the different waves of (30) shown as directions of propagation in Figure 15-7f are also the rays corresponding to such points as P_3 in Figure 15-2p(iii).

In the region of the geometrical shadow the total wave U_T is the sum of $U_i = e^{ikx}$ and U_+:

(31) $$U_{T+} = U_i + U_+ = -u_e(a - y) + u_e(a + y) \ ,$$

i.e., the shadow forming part of U_+ cancels the incident wave and we are left only with the edge waves or diffracted waves. Thus, corresponding to (31), a "perfect shadow" does not exist; the diffracted field is small for relatively small x (in the immediate vicinity of the obstacle), but it increases in magnitude as x increases and the shadow "disappears" with increasing distance from the scatterer. Although the present approximations are not adequate for either $x \sim 0$ or $x \sim \infty$, the results are qualitatively correct. The field U_{T+} is oscillatory both in x and y. The flux density along the axis $(y = 0)$, corresponding to $|U_{T+}|^2$ with

(32) $$U_{T+} \approx \sqrt{\frac{2x}{\pi k a^2}}\, e^{i\pi/4}\, e^{ikR} \ , \quad R = x + \frac{a^2}{2x} \approx \sqrt{x^2 + a^2} \ ,$$

is a relative maximum; this is the analog of the Arago "bright spot" discussed for the disk. In the back-scattered region inclosed by projections of the strip edges parallel to $-x$, we have

(33) $$U_{T-} = U_i + U_- = e^{ikx} + e^{-ikx} + u_e(a - y) + u_e(a + y)$$

where e^{-ikx} is the geometrically reflected wave.

The above results (24) to (33) are subject to two restrictions: the first, $\left|\frac{y - a}{x}\right| \ll 1$, enables us to use the approximations in (5), and restricts us to observation near the back and forward scattered directions: the second

$\frac{k(a \pm y)^2}{|x|} \gg 1$ is required in order to use approximation (22) for $\mathcal{F}$ and this restricts us to relatively moderate values of $|x|$. Together, the two restrictions state that $ka = \frac{2\pi a}{\lambda} \gg 1$ (i.e., the strip is very wide compared to the incident wavelength), and that y cannot be near $\pm a$ (i.e., y cannot approach the shadow boundaries).

We cannot use (31) to (39) for a point $y \sim a$ in the range of y_2 of Figure 15-7e. In that region although $-\eta_- = \sqrt{\frac{k}{\pi|x|}}\,(a + y) \approx \sqrt{\frac{k}{\pi|x|}}\,2a \gg 1$, we see that $\eta_+ = \sqrt{\frac{k}{\pi x}}\,(a - y) \sim 0$. Thus, in the general form (21), we still use (22) for $\mathcal{F}(-\eta_-)$, but we must use (23) for $\mathcal{F}(\eta_+)$. Consequently at the four points $|x|$, $|y_1|$, we obtain

$$(34) \qquad U_\pm = \mp e^{ik|x|}\left[\sqrt{\frac{k}{2\pi|x|}}\, e^{-i\pi/4}(a - y) + \frac{1}{2} - \sqrt{\frac{|x|}{2\pi k}}\, e^{i\pi/4}\, \frac{e^{ik2a^2/|x|}}{2a}\right].$$

The second and third terms are of the form considered in (24). The first is a cylindrical wave, i.e., the wave of a line source on the edge -- a "true" edge wave decreasing as $\frac{1}{\sqrt{x}}$ with increasing distance. The third term becomes negligible for very large ka, for which case the total field for $x > 0$ reduces to

$$(35) \quad U_{T+} = U_i + U_+ \approx \frac{1}{2}\, e^{ikx} - H(kx)\frac{k(a - y)}{2}, \quad H(kx \equiv \left[e^{ikx}\sqrt{\frac{1}{2\pi kx}}\, e^{-i\pi/4}\right].$$

Thus the field in the neighborhood of the shadow-line is half the incident wave plus a cylindrical wave corresponding to a line source with source strength proportional to $k(a - y)$. As y approaches a, we see that U_T approaches $\frac{1}{2}U_i$ linearly; this holds whether y approaches a from above from "above" or "below" in Figure 15-7e. Similarly for $x < 0$, we have

$$(36) \qquad U_{T-} \approx U_i + U_- \approx e^{ikx} + \frac{1}{2}\, e^{-ikx} + H(k|x|)\frac{k(a - y)}{2}.$$

In the range of y_3 in Figure 15-7e, we have $-\eta_\pm \gg 1$, and we again use (22) for both Fresnel integrals in (21). However, in contrast to (24), the scattered wave at the four points $(|x|, |y_3|)$ is given by

$$(37)\qquad U_{\pm} = \mp \frac{e^{ik|x|}}{\sqrt{2}} e^{-i\pi/4}\left[\mathcal{F}(-\eta_+) + \mathcal{F}(-\eta_-)\right]$$

$$\approx \mp e^{ik|x|} \sqrt{\frac{|x|}{2\pi k}}\, e^{i\pi/4} + \left[\frac{e^{ik(y-a)^2/2|x|}}{y - a} - \frac{e^{ik(y+a)^2/2|x|}}{y + a}\right]$$

$$= \mp [u_e(a - y) + u_e(a + y)] ,$$

so that such points receive only the edge contributions of (30).

The above explicit approximations suffice for present purposes. A more complete discussion of the problem of the strip is given in introductory texts on optics in which the field at any point in space is usually computed graphically from Cornu's spiral Figure 15-7d.[*]

Thus up to moderately large $|x|$ the scattered wave is largely "confined" to the strip $|y| < a$ in the sense that within this strip we obtain the geometrically reflected and shadow forming waves. These two waves correspond to the waves scattered by an infinite plane; however, superimposed on these are the additional waves that we interpreted as edge waves. Because of the additional waves, the shadow is not "perfect"; careful observations (subject to the present restrictions on distance parameters) on the shadows of scatterers having very regular edges show a system of bright and dark bands parallel to the edges of the scatterer (a "fringe system"). We now determine the number and separation of such extrema that may be observed in the shadow region on a screen parallel to the strip.

We use

$$(38)\qquad \frac{d}{dx}\int_0^{\phi(x)} F(\eta)\,d\eta = F(\phi(x))\phi'(x) ,$$

(compare Exercises 7-2, No. 3(a)), to differentiate U as given by (15) and (16), and obtain

$$(39)\qquad \frac{dU}{dy} \propto \frac{d}{dy}[\mathcal{F}(\eta_+) - \mathcal{F}(\eta_-)] = \left(e^{ik\eta_+^{\,2}/2}\right)\frac{d\eta_+}{dy} - \left(e^{i\pi\eta_-^{\,2}/2}\right)\frac{d\eta_-}{dy} ;$$

$$\frac{d\eta_+}{dy} = \frac{d\eta_-}{dy} ,$$

---

[*] See for example, F.A. Jenkins and H.E. White, Fundamentals of Optics, McGraw-Hill, 1957; ch. 18.

(compare Exercises 7-2, No. 3(b)). The extrema of U, obtained from $\frac{dU}{dy} = 0$, correspond to

(40) $$e^{i\pi\eta_+^2/2} = e^{i\pi\eta_-^2/2} .$$

Consequently the exponents must satisfy

(41) $$\frac{\pi}{2}(\eta_+^2 - \eta_-^2) = -\frac{2k}{x}ay = 2n\pi ; \quad n = 0 , \pm 1 \dots .$$

The distance between successive extrema is thus

(42) $$|y_{n+1} - y_n| = \Delta y = \frac{\pi x}{ka}$$

and there are N extrema, with N given by

(43) $$N = \frac{2a}{\Delta y} = \frac{2ka^2}{\pi x} = \frac{4a^2}{\lambda x} ,$$

in the geometrical projection of the slit. Thus the number of extrema increases with increasing strip width, or with decreasing wavelength, or decreasing axial distance.

Before continuing the main line of this section, we consider the range of very large $|x|$ excluded in the discussion based on (22). If $|x|$ becomes very large, so that $\frac{k(a \pm y)^2}{|x|} \approx 0$, then both $\eta_+ \approx 0$ and $\eta_- \approx 0$ in (21), and we approximate both Fresnel integrals by means of (14). Thus

(44) $$U_\pm \approx \frac{e^{-i\pi/4}}{\sqrt{2}} e^{ik|x|} \left[\sqrt{\frac{k}{\pi|x|}}\right] [a - y + a + y]$$

$$= \mp \left\{ \sqrt{\frac{2}{\pi k|x|}} e^{-i\pi/4} e^{ik|x|} \right\} ka = \mp H(k|x|)ka .$$

Thus for this case, the scattered field is essentially that of a line source of strength ka. We derived this result via (5) which means that it is restricted to angles near the forward and back directions; comparing with (4), we see that (43) is merely the special case of (4) corresponding to $\theta = 0$, π, and that for other values of θ, the appropriate form of U at large distances is simply (4).

Method of Stationary Phase. We have discussed the preliminaries, and can now turn to the main topic of this section. In general, we consider an integral of the form

(45) $$I = \int_{-a_-}^{a_+} G(x)e^{ikL(x)}dx \ , \qquad k \gg 1 \ ,$$

where G is a slowly varying function of x compared to $e^{ikL(x)}$ in the sense that fractional change in G is small when kL changes by 2π. If there exist one or more values of x for which $\frac{dL}{dx}$ vanishes, then the principal contributions to the value of the integral arise from the neighborhoods of the extrema (or stationary values) of L; elsewhere the contributions cancel through destructive interference as defined previously. We reiterate that the intuitive basis of this idea is the recognition that on an Argand diagram (as in Figures 2, 3, and 4), I is a sum of elementary vectors whose direction (essentially the phase kL) is in general a rapidly changing function of x, so that the resultant I is consequently small. However, if there exists a value of x for which $\frac{dL}{dx}$ vanishes, then the phase is stationary at this value and only slowly varying in its vicinity; the elementary vectors near this value are almost in phase and add to give a large result.

Analytically, we proceed as follows. If a value x_s exists for which $L'(x_s) = 0$, then the second order Taylor polynomial for L at x_s is

(46) $$L(x) = L(x_s) + L''(x_s)\frac{(x - x_s)^2}{2} + \dots .$$

Assuming that $L''(x_s) \neq 0$, and that the higher order terms are negligible, we keep only terms up to second order in the exponent of (45). We replace the slowly varying function $G(x)$ by its value $G(x_s)$ at the stationary point (the point marking the center of the region in which the integrand contributes significantly), and if x_s is the only stationary point we use the approximation

(47) $$I \approx I_s = G(x_s)e^{ikL(x_s)} \int_{-a_-}^{a_+} e^{ikL''(x_s)(x-x_s)^2/2} dx$$

$$= G(x_s)e^{ikL(x_s)} \sqrt{\frac{\pi}{kL''(x_s)}} \left[\mathcal{F}(\eta_+) - \mathcal{F}(\eta_-)\right] ,$$

$$\eta_\pm = \sqrt{\frac{kL''(x_s)}{\pi}} \, (\pm a_\pm - x_s) ,$$

where $\mathcal{F}$ is the Fresnel integral as in (10)ff. (If there is more than one stationary point, then I_s is a sum of such expressions.)

In particular, if $\pm\eta_{\pm} \sim \infty$, then $\mathcal{F} \sim \mathcal{F}(\infty)$ of (13); we have $\mathcal{F}(\infty) - \mathcal{F}(-\infty) = \sqrt{2}\, e^{-i\pi/4}$, and consequently

$$(48) \qquad I_s \sim G_s e^{ikL_s + i\pi/4} \sqrt{\frac{2\pi}{kL''_s}} ,$$

where the subscript s indicates that the function is evaluated at the stationary point x_s . Compare (48) with (7).

Before applying (48), let us show that in the small wavelength limit, $k \gg 1$, I_s is much larger than an integral I as given in (45) when the integrand in (45) has no stationary point. We show that if there are no stationary values of L , then the integral (45) only has the order of $\frac{1}{k}$ as compared to I_s which is itself proportional to $\frac{1}{\sqrt{k}}$; since $k \gg 1$, I_s is therefore much larger. To show this we introduce L as the new integration variable:

$$(49) \qquad I = \int_{-a_-}^{a_+} G(x) e^{ikL(x)} dx = \int_{L(-a_-)}^{L(a_+)} \frac{G(L)}{L'} e^{ikL} dL$$

$$= \frac{1}{ik} \int_{L(-a_-)}^{L(a_+)} \left[\frac{G(L)}{L'}\right] d(e^{ikL}) ,$$

and integrate by parts in order to develop I in powers of $\frac{1}{k}$:

$$(50) \qquad I = \frac{1}{ik} \left[\frac{G(L)}{L'} e^{ikL}\right]_{L(-a_-)}^{L(a_+)} + \frac{1}{k^2}\left[\frac{d}{dL}\left(\frac{G(L)}{L'}\right) e^{ikL}\right]_{L(-a_-)}^{L(a_+)} + \dots .$$

Thus as long as $\frac{L'}{G}$ does not vanish in the range $-a_-$ to a_+ , the integral is only of order $\frac{1}{k}$ and is therefore much smaller than the stationary case I_s of (48).

As a first application, we apply (47) to the original integral (3) for the strip with the constant given by (20):

$$(51) \qquad U = g \sqrt{\frac{k}{2\pi}}\, e^{-i\pi/4} I , \quad I = \int_{-a}^{a} \frac{e^{ikR}}{\sqrt{R}} d\eta , \quad R = \sqrt{x^2 + (y - \eta)^2} .$$

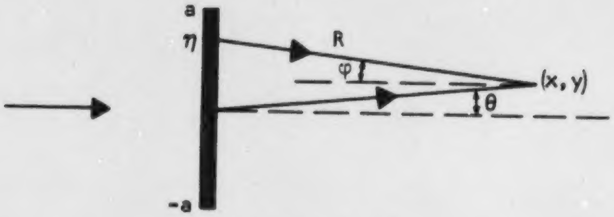


Figure 15-7g

Comparing with (45), we see that $G(\eta) = \dfrac{1}{\sqrt{R(\eta)}}$, and that $L(\eta) = R(\eta)$. Introducing ϕ as in Figure 15-7g, we differentiate to obtain

$$\frac{dL}{d\eta} = L' = R' = \frac{\eta - y}{\sqrt{x^2 + (\eta - y)^2}} = \frac{\eta - y}{R} = \sin\phi \,, \tag{52}$$

$$R'' = \frac{1}{\sqrt{x^2 + (y - \eta)^2}} - \frac{(\eta - y)^2}{[x^2 + (y - \eta)^2]^{3/2}} = \frac{1}{R}\left(1 - \frac{(\eta - y)^2}{R^2}\right) = \frac{\cos^2\phi}{R} \,. \tag{53}$$

The stationary values correspond to $R' = 0$:

$$\eta_s = y \;;\; \phi_s = 0 \,,\, \pi \,. \tag{54}$$

Consequently

$$R_s = \pm x = |x| \quad , \quad |R''_s| = \frac{\cos^2\phi}{R_s} = \frac{1}{R_s} = \frac{1}{|x|} \,, \tag{55}$$

substituting into the integral of (51) via (47), we obtain

$$I = \int_{-a}^{a} \frac{e^{ikR}}{\sqrt{R}}\, d\eta \approx \frac{1}{\sqrt{R_s}}\, e^{ikR_s} \sqrt{\frac{\pi}{kR''_s}} \Big[\mathscr{F}(\eta_+) - \mathscr{F}(\eta_-)\Big] \tag{56}$$

$$= \sqrt{\frac{\pi}{k}}\, e^{ik|x|} \Big[\mathscr{F}(\eta_+) - \mathscr{F}(\eta_-)\Big]$$

$$\eta_\pm = \sqrt{\frac{kR''_s}{\pi}}\,(\pm a - \eta_s) = \sqrt{\frac{k}{\pi|x|}}\,(\pm a - y) \,.$$

Introducing (56) into (51) yields the earlier form (21). The limiting form based on (48) gives $U = ge^{ik|x|}$ as previously, with $g = -1$ from our "shadow condition."

We may generalize the preceeding analysis directly to an arbitrary angle of incidence (see Figure 15-7h). The incident plane wave may be written

$$U_i = e^{ikx\cos\alpha + iky\sin\alpha} \,, \tag{57}$$

which is merely the form of (1) obtained by rotating the xy coordinate frame through an angle $-\alpha$. Since the phase of (57) is zero at the origin (the center of the strip), the wavelet originating at the origin has the same phase as obtained from (2). However, the wavelet originating at η is excited by

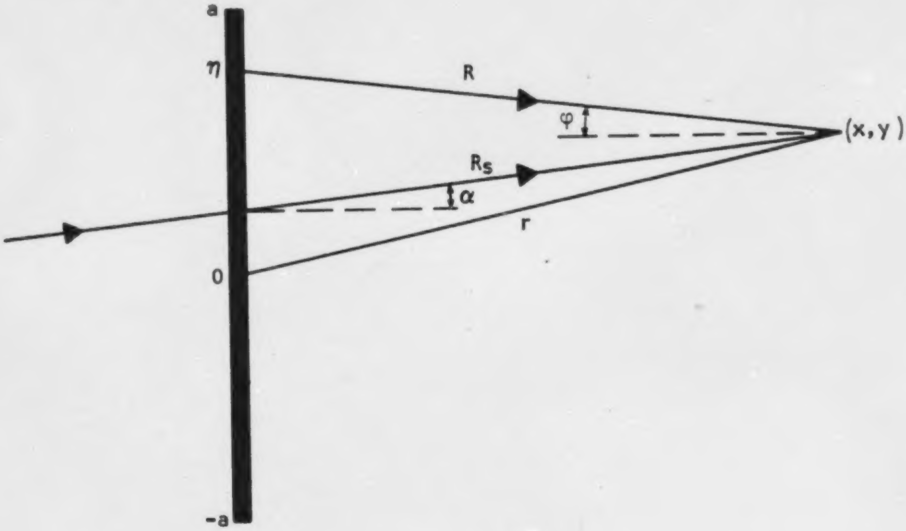


Figure 15-7h

$U_i(0,\eta) = e^{ik\eta \sin\alpha}$, so that its phase contains the additional term $r\eta \sin \alpha$. Thus instead of (2) we have

$$u(\eta) = c\,\frac{e^{ikR+ik\eta \sin \alpha}}{\sqrt{kR}}, \tag{58}$$

and corresponding to Figure 15-7h, we replace (51) by

$$U = \int_{-a}^{a} c\,\frac{e^{ikR+ik\eta \sin \alpha}}{\sqrt{kR}}\,d\eta, \tag{59}$$

where the appropriate value of c will be determined from a limiting case.

Corresponding to (45), we take $G = \frac{c}{\sqrt{kR}}$ to be slowly varying, and differentiate the phase

$$L = \eta \sin \alpha + \sqrt{x^2 + (\eta - y)^2}, \tag{60}$$

with respect to η. We now have

$$L' = \sin \alpha + \sin \phi, \quad L'' = \frac{\cos^2 \phi}{R}. \tag{61}$$

The stationary values correspond to

$$(62)\qquad L' = \sin\alpha + \sin\phi = 0 \ ; \quad \sin\phi = -\sin\alpha \ ; \quad \phi = -\alpha \ , \ \pi + \alpha \ ,$$

which contains Euclid's principle of reflection and the principle of shadow formation. Since $\sin\phi = \frac{\eta - y}{R} = \frac{\eta - y}{x}\cos\phi$, we see from (62) that $y - \eta_s = |x| \tan\alpha$, $R_s = x \sec\phi = |x| \sec\alpha$, and

$$(63)\ L_s = \eta_s \sin\alpha + x \sec\phi_s = (y - |x| \tan\alpha)\sin\alpha + |x|\sec\alpha = y \sin\alpha + |x|\cos\alpha \ .$$

Similarly,

$$(64)\qquad L''_s = \frac{\cos^2\alpha}{R_s} \ .$$

Substituting into (59) in terms of the limiting form (48) we obtain

$$(65)\qquad U \sim \frac{c_s}{\sqrt{kR_s}} e^{+ikL_s + i\pi/4} \frac{2\pi}{\sqrt{k|L''_s|}}$$

$$= e^{\pm ikx\cos\alpha + iky\sin\alpha} \frac{c_s}{k\cos\alpha} \sqrt{2\pi}\, e^{i\pi/4} = U_{\pm} \ .$$

Comparing U_+ with U_i of (57), we determine c_s from the shadow condition $U_+ = -U_i$:

$$(66)\qquad c_s = -\frac{k \cos}{\sqrt{2\pi}} e^{-i\pi/4} \ ,$$

which differs from our earlier result by the additional factor $\cos\alpha$. Similarly, the corresponding value of g for a perfect reflector as for (24)ff, is now replaced by $g \cos\alpha$.

Using the more general form corresponding to (47), we now have

$$(67)\qquad U = g e^{ik|x|\cos\alpha + iky\sin\alpha} e^{-i\pi/4} \frac{1}{\sqrt{2}} [\mathcal{F}(\eta_+) - \mathcal{F}(\eta_-)] \ .$$

$$\eta_{\pm} = \sqrt{\frac{k}{\pi|x| \sec\alpha}} \cos\alpha(\pm a + |x| \tan\alpha - y) \ .$$

Circular Cylinder. Let us now apply the same method to consider scattering of the plane wave (1) by the convex cylinder as in Figure 15-7i. The point $a(\phi)$ on the cylinder has the coordinates $a\cos\phi$, $a\sin\phi$, and the excitation at the point is $e^{ika\cos\phi}$. We write the corresponding scattered field as the integral of u over the arc $a d\phi$:

(68) $$U = \int_{\pi/2}^{3\pi/2} \frac{c}{\sqrt{kR}} e^{ik(R+a\cos\phi)} a d\phi .$$

We do not know c completely, but the result (66) and the corresponding form $g \cos\alpha$ suggest the generalization in which α (the angle of incidence with respect to the surface normal) is replaced by $x - \phi$. Thus for convenience (and as can be justified by a more complete model) we use

(69) $$U = -ga \sqrt{\frac{k}{2\pi}}\, e^{-i\pi/4}\, I \quad , \quad I = \int_{\pi/2}^{3\pi/2} \frac{e^{ik(R+a\cos\phi)}}{\sqrt{R}} \cos\phi \, d\phi ,$$

and we shall see that $\cos\phi$ in the integrand is appropriate for both geometrical reflection and shadow formation. We consider only the range $0 < \theta < \pi$ explicitly; however, the results may be extended to all θ by introducing absolute values of the trigonometric functions (as required to preserve symmetry).

The phase of I is proportional to

(70) $$L = R + a\cos\phi = \sqrt{r^2 + a^2 - 2ar\cos(\phi - \theta)} + a\cos\phi ,$$

and its derivative

(71) $$\frac{dL}{d\phi} = L' = \frac{ar\sin(\phi - \theta)}{R} - a\sin\phi ,$$

vanishes for the two values $\phi = \phi_L$ or ϕ_D, such that

(72L) $$\frac{\sin(\phi_L - \theta)}{R_L} = \frac{\sin\phi_L}{r_L} ,$$

(72D) $$\frac{\sin(\phi_D - \theta)}{R_D} = \frac{\sin(\pi - \phi_D)}{r_D} .$$

For a given value of θ, the phase has only one stationary value; the value may correspond either to geometrical reflection as in Figure 15-7j, or to forward scattering in Figure 15-7k. The first value applies for y in the "lit" region L in Figure 15-7j; the second value, which yields the shadow-forming rays, corresponds to y in the "dark" region D as in Figure 15-7k. Equation (72L) corresponds to Euclid's principle of reflection, and (72D) to the principle of shadow formation.

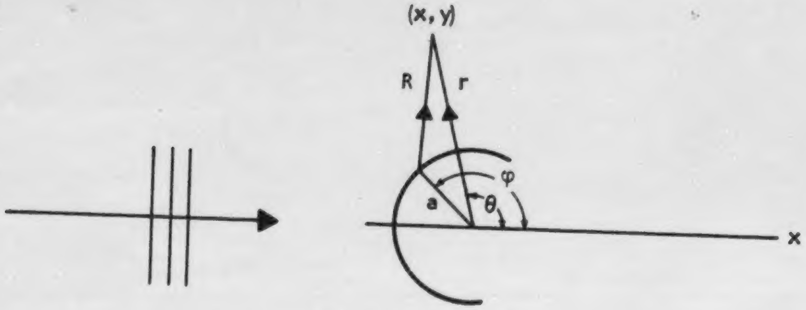


Figure 15-1

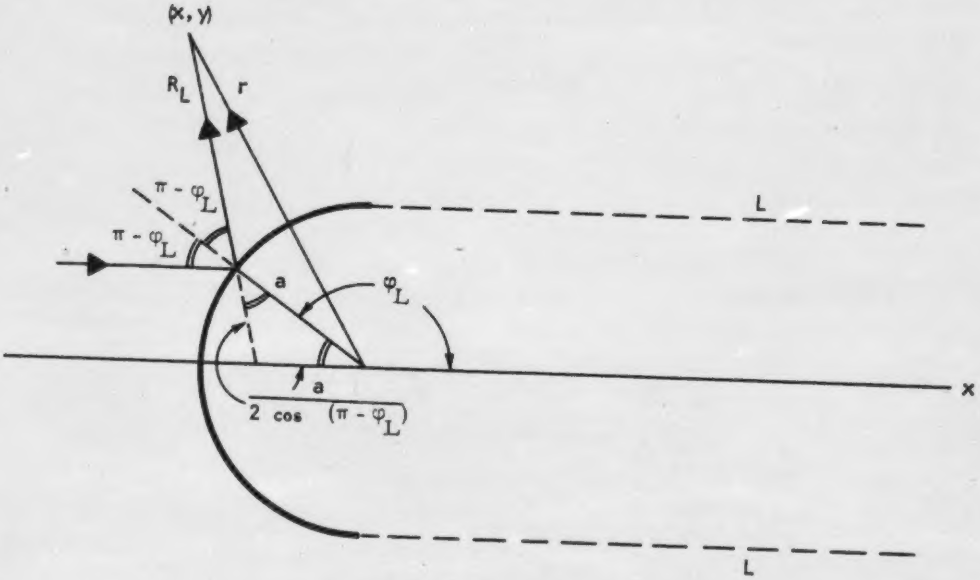

(x, y)
R_L
r
$\pi - \varphi_L$
$\pi - \varphi_L$
L
a
φ_L
$\frac{a}{2 \cos (\pi - \varphi_L)}$
x
L

Figure 15-7j

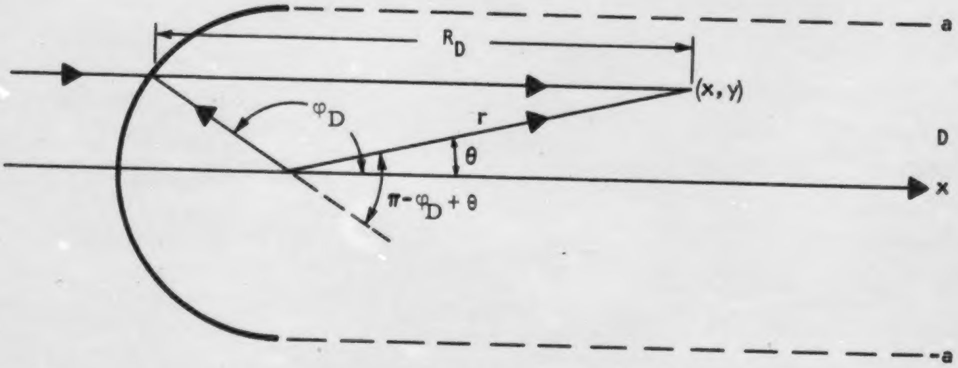

a
R_D
(x, y)
φ_D
r
D
θ
x
$\pi - \varphi_D + \theta$
-a

Figure 15-7k

The second derivative of L equals

$$\text{(73)} \qquad \frac{d^2L}{d\phi^2} = L'' = -\frac{a^2r^2 \sin^2(\phi - \theta)}{R^3} + \frac{ar}{R}\cos(\phi - \theta) - a\cos\phi ,$$

and substituting (72) leads to the special values at the stationary points. Thus for either case,

$$\text{(74)} \qquad L''_s = \frac{a}{R_s}[-a \sin^2 \phi_s + r\cos(\phi_s - \theta) - R_s \cos \phi_s] ,$$

where R_s and ϕ_s are the special values shown in either Figure 15-7j or Figure 15-7k. At a forward point, we see from Figure 15-7k that
$r\cos(\pi - \phi_D + \theta) + a = R_D \cos(\pi - \phi_D)$; consequently,
$r\cos(\phi_D - \theta) - R_D \cos\phi_D = a$, and (61) reduces to

$$\text{(75D)} \qquad L''_D = \frac{a^2}{R_D}\cos^2 \phi_D .$$

On the other hand, at the reflection point, we see from Figure 15-7j that
$r\cos(\phi_L - \theta) = a + R_L\cos(\pi - \phi_L) = a - R_L \cos \phi_L$; consequently

$$\text{(75L)} \qquad \begin{aligned} L''_L &= \frac{a}{R_L}[a(1 - \sin^2 \phi_L) - 2R_L \cos \phi_L] \\ &= \frac{2a\cos(\pi - \phi_L)}{R_L}[R_L + \frac{a}{2}\cos(\pi - \phi_L)] . \end{aligned}$$

Similarly the stationary value of the phase function at a forward point equals

$$\text{(76D)} \qquad L_D = R_D + a\cos\phi_D = R_D - a\cos(\pi - \phi_D) = x .$$

Although we could eliminate R_L and ϕ_L in the expression for the phase of the reflected wave, it is simpler to leave L_L in the original form

$$\text{(76L)} \qquad L_L = R_L - a\cos(\pi - \phi_L) ;$$

here (for a given value of r and θ) R_L and ϕ_L are determined as in Figure 15-7j, i.e., by noting which ray (determined by the value of y) of the incident wave front can reach $r(\theta)$ via reflection in the cylindrical surface.

Finally the required slowly varying parts of the integrand if I evaluated at the stationary points are given by

(77) $$G_s = \frac{\cos \phi_s}{\sqrt{R_s}} \quad ; \quad s = D\ ,\ L\ .$$

Substituting (48) into (69), we obtain

(78) $$U_s = -g\ aG_s \frac{e^{ikL_s}}{\sqrt{L''_s}}\ .$$

Thus for the perfect reflector $g = 1$, we enter the above D-values in (78) and obtain the shadow forming wave

(79) $$U_D = -e^{+ikx} = -U_i\ .$$

Similarly, for the geometrical reflected wave (in the "lit" region L)

(80) $$U_L = g \sqrt{\frac{a \cos \alpha}{2(R_L + \frac{a}{2} \cos \alpha}}\ e^{ik(R_L - a \cos \alpha)} \quad , \quad \alpha = \pi - \phi_L\ .$$

The result (80) corresponds to the geometrically obtained results of Sections 15-2 and Section 15-4. In particular, the flux density ratio $\frac{F}{F_0}$ of (14) of Section 15-4 is simply the present $|U_L|^2$, and similarly the (18) of Section 15-4 corresponds to $|U_s|^2$ of (78); the present forms are richer in that they make the roles of the ray path (L_s) and the caustic (L''_s) explicit.

The corresponding Fresnel approximations are obtained by using (47) instead of (48). We change variables to $\eta = a \sin \phi$, $D_\phi L = a \cos \phi\ D_\eta L$, and obtain

(81) $$\mathcal{U}_D = U_D \frac{e^{-i\pi/4}}{\sqrt{2}} [\mathcal{F}(\eta_+) - \mathcal{F}(\eta_-)] \quad , \quad \eta_\pm = \sqrt{\frac{k}{\pi x}}\,(\pm a - y)\ ,$$

where U_D is given in (79); thus (81) is simply (21) for the range $x > 0$. Similarly, in terms of U_L of (80), we have

(82) $$\mathcal{U}_L = U_L \frac{e^{-i\pi/4}}{\sqrt{2}} [\mathcal{F}(\eta_+) - \mathcal{F}(\eta_-)]\ ,$$

$$\eta_\pm = \sqrt{\frac{2k(R_L + \frac{a}{2} \cos \alpha)}{\pi R_L\ a \cos \alpha}}\,(\pm a - a \sin \alpha) \quad , \quad \alpha = \pi - \phi_L\ ,$$

where $\mathcal{F}(\eta)$ is the Fresnel integral as in (10)ff.

The **Field On Caustics**. Although supplemented by phase considerations, Equations (79) and (80) are still results of geometrical optics which we could construct piece-by-piece from the special "laws" of the earlier sections:

> [H'] gives the directions, [KL] the magnitudes, and the phases may be obtained from Newton's idea of periodicity.

However we have now obtained these results essentially from the single idea of periodic waves that evolved through the work of Huyghens, Young, and Fresnel into the integral (3). Starting with (3), we used a mathematical procedure to approximate the integral and obtain the limiting form (48) which converts (3) to the general short-wavelength approximation (7) (two of whose special cases are given by (78) - (80)). Not only does the integral (3) supercede the earlier special laws, but it governs many other phenomena than those covered by (7) or (78). We have already applied it to obtain the Fraunhofer and Fresnel approximations. We now apply it to determine the magnitude of the field on a caustic, the case $L''_S = L''_H = 0$ excluded in (46) and in our discussion of [KL] . Were we considering the analogous problem of a concave semicircle, then we could pick r to satisfy $L''_S = 0$, so that the denominator of (80) would vanish. However, this would not require a special "law" to set right -- merely a more complete approximation of (3) than given by (48).

On a caustic, both L' and L'' vanish; in addition, on a cusp of a caustic (where the derivative of the equation of the caustic vanishes, since the curve changes direction) the third derivative L''' must also vanish. Thus for such cases we can no longer approximate $L(x)$ by means of (46), i.e., we must keep additional terms in the Taylor approximation in order to obtain the first correction to $L(x_s)$.

Let us assume that the first $n - 1$ derivatives of L at x_s vanish. We then approximate $L(x)$ by the Taylor polynomial of n-th order

$$(83) \quad L(x) = L(x_s) + \frac{(x - x_s)^n}{n!}\left(\frac{d^n}{dx^n} L(x)\right)_{x=x_s} = L_s + (x - x_s)^n \frac{L_s^{(n)}}{n!} .$$

In this case, for infinite ends of integration (corresponding essentially to (48), we have

$$(84) \qquad I = \int_{-\infty}^{\infty} G(x)e^{ikL(x)}dx \approx G_s e^{ikL_s} \int_{-\infty}^{\infty} e^{ik(x-x_s)^n F^{(n)}/n!} dx .$$

If we introduce a new variable y through $y^n = \frac{k(x - x_s)^n L_s^{(n)}}{n!}$, then we may rewrite (84) as

$$I \approx G_s e^{ikL_s} \left(\frac{n!}{kL_s^{(n)}} \right)^{1/n} \int_{-\infty}^{\infty} e^{iy^n} \, dy , \tag{85}$$

where the remaining integral depends only on n.

Using (85) instead of (48) in (68), we obtain

$$U_L \approx C_n e^{ikL_s} (k)^{1/2-1/n} \tag{86}$$

where we have suppressed practically everything but the dependence on $k = \frac{2\pi}{\lambda}$. Thus for the line caustic of the circular cylinder, we have $n = 3$, and since $R(\underset{\sim}{r}') \approx a(\underset{\sim}{r}')$ on the caustic $R = \frac{a}{2} \cos \alpha$, we see from (9) and from dimensional considerations that U_L is proportional to $(ka)^{1/6}$; similarly for a cusp ($\alpha = 0$ for the circle), we have $n = 4$, and U_L is proportional to $(ka)^{1/4}$. Thus U_L on the caustic and focus increases as the wavelength λ decreases, and indeed would be infinite if there were such a thing as a zero wavelength (the implicit assumption of conventional geometrical optics). More generally, since $n \geq 3$, we see that $k^{1/2-1/n}$ always increases as λ decreases.

Exercises 15-7

1.. Sketch vector diagrams for the first zero and first and second extrema of $\Gamma(\theta) = \frac{\sin(ka \sin \theta)}{ka \sin \theta}$ of Equation (4), where $\theta \geq 0$.

15-8. Mathematical Model for Scattering.

In previous sections we considered certain aspects of wave theory but based the development on several unrelated "laws of nature." In the present section we tie these special postulates together into a mathematical model for scattering.

Wave Equation. We considered the plane waves $\cos(\pm kx - \omega t)$, and, for convenience, worked with the real part of the corresponding exponentials:

$$(1) \qquad e^{\pm ikx-i\omega t} = e^{\pm ikx} e^{-i\omega t} = f(x)g(t) .$$

Equation (1) is a product of a function of x times a function of t , each of which satisfies a second order differential equation:

$$(2) \qquad \frac{d^2f(x)}{dx^2} = -k^2f(x) \quad , \quad k = \frac{2\pi}{\lambda} ;$$

$$(3) \qquad \frac{d^2g(t)}{dt^2} = -\omega^2g(t) = -k^2v^2g(t) \quad , \quad \omega = kv = \frac{2\pi v}{\lambda} = 2\pi\nu .$$

As discussed previously, v is the phase velocity of the wave generated by a source vibrating at a frequency ν , and λ is the wavelength -- the distance between crests.

The general form of (2) and (3) is

$$(4) \qquad \frac{d^2F(y)}{dy^2} + \beta^2F(y) = 0 ,$$

whose solutions equal

$$(5) \qquad \cos\beta y \quad , \quad \sin\beta y \quad , \quad e^{i\beta y} \quad , \quad e^{-i\beta y} \quad ,$$

or any linear combination of these functions. Thus in choosing the particular combinations that led to (1), we used some selection rules. We discuss these rules subsequently.

Now let us use the above to construct more general equations. Our attitude is the following: We know of phenomena that can be described by wave functions such as (1). Let us seek a general wave equation that yields (1) as well as more general wave forms. The more general waves may well correspond to phenomena not covered by (1).

From (2) and (3), we have

$$\frac{1}{f(x)}\frac{d^2f(x)}{dx^2} = -k^2 \quad , \quad \frac{1}{v^2g(t)}\frac{d^2g(t)}{dt^2} = -k^2 \ .$$

Subtracting one from the other, we obtain

$$(6) \qquad \frac{1}{f(x)}\frac{d^2f(x)}{dx^2} - \frac{1}{v^2g(t)}\frac{d^2g(t)}{dt^2} = 0 \ ,$$

or equivalently

$$(7) \qquad g(t)\frac{d^2f(x)}{dx^2} - \frac{f(x)}{v^2}\frac{d^2g(t)}{dt^2} = 0 \ ,$$

where $\frac{d}{dx}$ operates only on $f(x)$, and $\frac{d}{dt}$ on $g(t)$. The present notation is awkward. We would like to combine $f(x)g(t)$ in a single form $E(x,t)$. To do so and preserve the idea that the differentiations with respect to x and t are independent, we introduce the notation $\frac{\partial}{\partial x} = D_x$ to represent differentiation with respect to x while t is fixed -- partial differentiation; similarly for t, $\frac{\partial}{\partial t} = D_t$. Thus, we rewrite (7) as

$$(8) \qquad \frac{\partial^2E(x,t)}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2E(x,t)}{\partial t^2} = \left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{v^2\partial t^2}\right)E(x,t) = 0 \ .$$

This is called the wave equation. The wave functions of (1) are special solutions of (8) corresponding to periodic waves.

We generalize (8) to two spatial dimensions x, y by introducing an additional operation $\frac{\partial^2}{\partial y^2}$ into (8):

$$(9) \qquad \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} - \frac{\partial^2}{v^2\partial t^2}\right)E(x,y,t) = 0 \ .$$

The plane waves $e^{\pm kx\cos\alpha + iky\sin\alpha - i\omega t}$ that we considered in Section 15-7 are solutions of (9) (see Exercises 15-8, No. 1). Similarly for three spatial dimensions we introduce an additional operation $\frac{\partial^2}{\partial z^2}$ in (9). We write the general form as

$$\left(\Delta - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) E(\vec{r},t) = \Delta E(\vec{r},t) - \frac{1}{v^2}\frac{\partial^2 E(\vec{r},t)}{\partial t^2} = 0 , \tag{10}$$

where $\vec{r} = (x,y,z)$, and where

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{11}$$

(which is also frequently written ∇^2) is called Laplace's operator.

The elementary spherical wave that we considered previously is the special solution of (10) that depends only on the magnitude r of $\vec{r}$ and not on direction. The simpler equation for the elementary spherical wave

$$E(r,t) = \frac{e^{ikr-i\omega t}}{r} \tag{12}$$

can be obtained by comparison with (1) and (8). Thus if we replace $E(x,t)$ by $rE(r,t)$ and $\frac{\partial^2}{\partial x^2}$ by $\frac{\partial^2}{\partial r^2}$, we obtain the corresponding equation for (12):

$$\left(\frac{\partial^2}{\partial r^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) rE(r,t) = 0 . \tag{13}$$

We may rewrite this directly in the form of (10):

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial E}{\partial r}\right) - \frac{1}{v^2}\frac{\partial^2 E}{\partial t^2} = 0 . \tag{14}$$

The general solution of (8) may be written

$$E(x,t) = F(vt - x) + G(vt + x) \tag{15}$$

where F and G are arbitrary. Similarly, the general solution of (13) is

$$rE(r,t) = F(vt - r) + G(vt + r) . \tag{16}$$

The corresponding solution for the equation of the line source in two dimensions and of the general equation (10) cannot be expressed so simply. We mention the general solutions only to stress that the solutions corresponding to periodic waves are special cases.

Let us now ignore practically everything that led us to the wave equation (10). We accept (10) as fundamental and seek its periodic solutions. For completeness, we repeat the definitions of the fundamental parameters given in previous sections.

The periodic waves we considered correspond to solutions having the product form

$$(17)\qquad E(\vec{r},t) = f(\vec{r})g(t) \ .$$

If we substitute (17) into (10), the "variables separate" in the sense that we obtain

$$(18)\qquad \frac{1}{f(\vec{r})}\,\Delta f(\vec{r}) = \frac{1}{v^2 g(t)}\,\frac{d^2 g(t)}{dt^2}$$

essentially as in (6) (see Exercises 15-8, No. 3). Since the left side of (18) is a function only of $\vec{r}$, and the right side only of t , each side must equal the same constant; call this constant $-k^2$. Thus (18) reduces to

$$(19)\qquad \frac{d^2 g(t)}{dt^2} + k^2 v^2 g(t) = 0 \ ,$$

$$(20)\qquad \Delta f(\vec{r}) + k^2 f(\vec{r}) = 0 \ ,$$

where (20) is known as Helmholtz's equation.

Equation (19) is the form (4) we considered previously. Its solutions are the periodic functions of (5). Without any loss of generality, we pick

$$(21)\qquad g(t) = e^{-ikvt} = e^{-i\omega t} \ , \quad \omega = kv$$

to work with. In Equation (10), v is given as the velocity: the distance an element of the wave covers in unit time. From (21), we see that $g(t)$ is periodic in t , i.e., if the time t changes by multiples of the constant $T = \frac{2\pi}{\omega}$ then g is unaltered:

$$(22)\qquad g(t) = g(t + mT) \quad ; \quad T = \frac{2\pi}{\omega} = \frac{1}{\nu} \ , \quad n = 1 \ , 2 \ , 3 \ , \ldots \ .$$

Thus T is the periodicity of the wave in time, and ν (the frequency) is the number of times that g has the same value in unit interval of time. The period in space, the wavelength $\lambda = vT = \frac{2\pi v}{\omega}$ is the distance covered by an element moving with velocity v for a time T . But from (21), we have $\frac{v}{\omega} = \frac{1}{k}$. Consequently $k = \frac{2\pi}{\lambda}$ is the relation between the "separation constant" (the propagation factor or wave number) and wavelength.

The space equation (20) is known as the reduced wave equation or Helmholtz's equation. The one-dimensional case $\frac{d^2 f(x)}{dx^2} + k^2 f(x) = 0$ is given in (2), and the special case of the spherically symmetrical wave is implicit in (14); i.e., $\frac{d^2}{dr^2}(rf) + k^2 rf = 0$, $f = f(r)$.

The above equations specify propagation of waves in a medium whose properties are determined solely by v. For the periodic cases, once we fix the frequency factor ω, the corresponding wavelength in the medium is determined. If we are dealing with several such media specified by different velocities v_m ; $m = 0, 1, \ldots$, then we obtain the same wave equations with v replaced by v_m ; the corresponding reduced wave equations for frequency factor ω involve $k_m = \frac{\omega}{v_m} = \frac{2\pi}{\lambda_m}$. For convenience we take v_0 as a reference, and write

$$v_m = \frac{v_0}{\mu} \tag{23}$$

where μ is the relative index of refraction. Consequently $k_m = \mu k$, and the corresponding space equation is

$$(\Delta + \mu^2 k^2) f(\vec{r}) = 0 , \tag{24}$$

e.g.,

$$\left(\frac{d^2}{dx^2} + \mu^2 k^2\right) f(x) = 0 \tag{25}$$

for the one-dimensional case. If μ is independent of x, then the solutions of (25) are the forms (5) with $\beta = \mu k$.

Conditions on the Solution. All the problems we considered are described by functions $E(\vec{r})g(t) = E(\vec{r})e^{-i\omega t}$, where $E(\vec{r})$ is a particular solution of the reduced wave equation

[I] : $$(\Delta + \mu^2 k^2) E(\vec{r}) = 0 .$$

The particular solution is determined by constraints that have been implicit in our development. The constraints are of two kinds.

[II] : restrictions on the solutions at the scatterer's surface,

[III]: restrictions on the solution at large distances from the scatterer.

The additional constraints are necessary because the wave equation merely describes the local properties of the medium and how a wave travels from point to point. But what if the medium is discontinuous? For example,

a. suppose we have a glass of water and consider waves on the surface of the water bounded by the unyielding rim of the glass;

b. suppose we are in a boat on a very large lake and the boat is an obstacle for an incoming wave.

Case a and b illustrate two essentially different kinds of wave problems with which we may be concerned.

In Case a we deal with a bounded medium: we are given v , the shape of the boundary and constraints on the solution at the boundary, and then may seek to determine the forms and periods of the waves that can be maintained in such enclosed media. These are free vibration problems: the waves on a taut clothesline, the waves on the surface of a glass of water, the sound waves in a closed room, the electromagnetic waves in a metal cavity; these are illustrations, and analogous problems exist in the quantum theory of atomic states.

If the bounding surface is one that yields, then waves on the inside create waves on the bounding surface, and they may propagate in a region external to the surface. We may also wet up vibrations on a surface and use the surface as a source of waves for the external medium, e.g., a vibrating drumhead as a source of sound. All musical instruments, strings, drums, pipes are examples of "vibrator-radiator" systems for sound. (We have switched from talking about light to talking about sound and water waves; this is at once for convenience and to stress the fact that for wave physics there are analogous phenomena in all branches of science.)

In Case b we deal with a bounded object that represents an obstacle to a wave traveling in an essentially unbounded medium. We require conditions that tell us the shape and size of the obstacle, whether its surface is penetrable by waves, and, if so, then what is the medium inside its surface. Such boundary conditions or transition conditions specify the kind of discontinuity the obstacle represents in the imbedding medium. Depending on the phenomena we seek to model, we may require boundary conditions such as

[IIa] $$E = 0 \text{ on surface}$$

or

[IIb] $$\frac{\partial E}{\partial n} = 0 \text{ on surface}$$

where $\frac{\partial E}{\partial n}$ is the rate of change of E along a normal at a point on the surface. These conditions correspond to surfaces impenetrable to waves. If the surface is penetrable (partially transparent), then many phenomena correspond to the following: the waves outside the scatterer's surface travel in medium-1 and the wave functions satisfy $(\Delta + \mu_1^2 k^2)E_1 = 0$; within the scatterer they travel in medium-2 and satisfy $(\Delta + \mu_2^2 k^2)E_2 = 0$; at the surface E_1 and E_2 are related by the transition conditions

[IIc] $$E_1 = E_2 \; , \quad \frac{\partial E_1}{\partial n} = A \frac{\partial E_2}{\partial n} \; ,$$

where A is a supplementary physical constant. Thus in general, the wave problems we consider are specified by two physical constants (or "physical parameters") μ and A whose values must be assigned at the start.

Having [I] and [II] , we complete the mathematical statement of the scattering problem by conditions at large distances from the scatterer [III] . These specify that we seek a solution consisting of essentially two terms: one term corresponds to the incident field, e.g., a plane wave

[IIIa] $$E_i = e^{ikx} \; ,$$

which is the space part of $e^{ikx-i\omega t}$; the other term, say E_s , corresponds to the wave outgoing from the obstacle that E_i excites. In three dimensions, we considered simple sources producing outgoing waves proportional to $\frac{e^{ikr-i\omega t}}{r}$ so that the wave surfaces were spheres of constant r . We also saw when we applied Huyghens' principle to geometrical reflection from a semicircle, that the wave surfaces that were complicated in shape in the vicinity of the obstacle became more and more symmetrical with increasing distance. We summarize all such cases by the statement

[IIIb] $$E_s \sim g \, \frac{e^{ikr}}{r} \quad \text{as} \quad r \sim \infty \; ,$$

where r is measured from some point in the scatterer, and where g (called the scattering amplitude) is independent of r . Thus at large distances from the scatterer $(r \sim \infty)$, E_s reduces to the elementary symmetrical wave times a function of angles. Similarly for a line scatterer, we use the elementary form $\frac{e^{ikr}}{\sqrt{r}}$ and write

[IIIc] $$E_s \sim g \frac{e^{ikr}}{\sqrt{r}} \quad \text{as} \quad r \sim \infty \; .$$

For on planar scatterer, the elementary source is $e^{ik|x|}$ and we have

[IIId] $$E_s \sim g e^{ik|x|} \quad \text{as} \quad |x| \sim \infty \; .$$

Collectively we write the total field as

[III'] $$E = E_i + E_s \; ; \; E_i = e^{ikx} \; ; \; E_{sm} \sim g_m \frac{e^{ikr}}{r^{(m-1)/2}} \; ; \; m = 1, 2, 3 \; .$$

where i and s stand for incident and scattered respectively.

Equations [I] , [II] , and [III] constitute the mathematical model for scattering. They replace all the special principles we considered previously; they cover all the cases where the principles apply, and many additional ones as well.

Point Scatterer. As an elementary illustration let us consider the scattering of a plane wave e^{ikx} by a small sphere of radius a for the boundary condition [IIa] $E = 0$ at $r = a$. For the general case of a sphere of arbitrary radius a, we would work with the complete solution of [I] for $\mu = 1$ subject to [IIIa,b,c] . We would represent E_a and E_s in terms of angle-dependent functions and initially unknown constants, and then use [IIa] to determine the constants. However, the restriction $a \ll \lambda$, or equivalently,

(26) $$ka \sim 0$$

simplifies the problem. From the geometry of Figure 15-8a, the incident field e^{ikx} equals $e^{ika \cos \phi}$ at the surface of the sphere; using the restriction (26), we have $e^{ika \cos \phi} \sim 1$, so that we may work with

(27) $$E_i(a) = 1 \; .$$

Thus the exciting field at the surface is independent of angles, and the corresponding scattered wave must be similarly independent of angles; $E_s(r)$ is a solution of $\frac{d^2}{dr^2}(rE_s) + k^2 rE_s = 0$, and the only one satisfying [IIIb] at large distances is

(28) $$E_s = C \frac{e^{ikr}}{r}$$

where C is a constant. At the surface of the scatterer $r = a$, we have $\frac{e^{ikr}}{r} \sim \frac{1}{a}$, and consequently the total field at $r = a$ is

(29) $$E(a) = E_i(a) + E_s(a) = 1 + \frac{C}{a} \; .$$

Applying the boundary condition $E(a) = 0$, we get

(30) $$C = -a \; .$$

The scattered wave for $r > a$ is thus

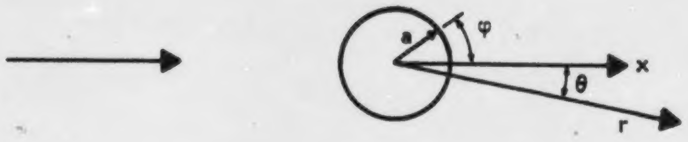


Figure 15-8a

$$E_s = -\frac{a}{r}e^{ikr} . \tag{31}$$

This corresponds physically, for example, to scattering of underwater sound by a small air-bubble.

Slab Scatterer. As another example, let us consider scattering of a plane wave by a partially transparent slab as in Figure 15-8b. The conditions on the problem are:

$$\left(\frac{d^2}{dx^2} + k^2\right)E_1 = 0 \ , \qquad |x| > 0 \ ; \tag{32}$$

$$\left(\frac{d^2}{dx^2} + K^2\right)E_2 = 0 \ , \quad K = \mu k \ ; \qquad |x| < 0 \ ; \tag{33}$$

$$E_1 = E_2 \ , \quad \frac{dE_1}{dn} = A\,\frac{dE_2}{dn} \ , \qquad |x| = a \ . \tag{34}$$

$$E_1 = E_i + E_s \ : \ E_i = e^{ikx} \ , \ E_s \sim g e^{ik|x|} \quad \text{as} \quad |x| \sim \infty \ . \tag{35}$$

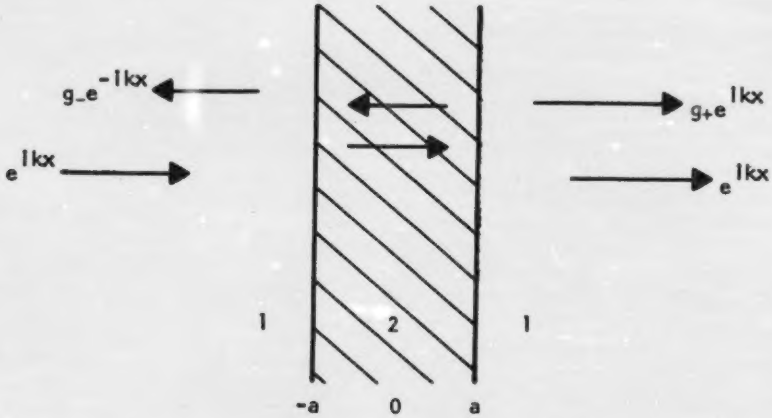


Figure 15-8b

From (35), we write

$$E_s = \begin{cases} g_+ e^{ikx} , & \text{for } x > a \\ g_- e^{ikx} , & \text{for } x < -a . \end{cases} \tag{36}$$

From (33) we take the most general solution in the form

$$E_2 = b_+ e^{iKx} + b_- e^{+iKx} . \tag{37}$$

We thus have four constants (g_+, g_-, b_+, b_-) to determine, and we do so by applying the surface conditions (34).

At $x = -a$, we get

$$e^{-ka} + g_- e^{ika} = b_+ e^{-iKa} + b_- e^{iKa} \tag{38}$$

$$k(e^{-ka} - g_- e^{ika}) = AK(b_+ e^{-iKa} - b_- e^{iKa}) ; \tag{39}$$

similarly, at $x = +a$,

$$b_+ e^{iKa} + b_- e^{-iKa} = (1 + g_+) e^{ika} \tag{40}$$

$$KA(b_+ e^{iKa} - b_- e^{-iKa}) = k(1 + g_+) e^{ika} . \tag{41}$$

Thus we have four algebraic equations for the four unknowns.

Solving (38) - (41) (Exercises 15-8, No. 4), and introducing the abbreviation

$$Q = \frac{Z - 1}{Z + 1} , \quad Z = \frac{KA}{k} = \mu A , \tag{42}$$

we obtain

$$g_- = -Q \frac{e^{-i2ka}(1 - e^{i4Ka})}{1 - Q^2 e^{i4Ka}} = R \tag{43}$$

$$g_+ + 1 = \frac{(1 - Q) e^{i(K-k)2a}}{1 - Q e^{2ika}} = T \tag{44}$$

where R and T are called the reflection and transmission coefficients. The corresponding internal field is, from (37),

$$E_2 = (1 - Q) e^{i(K-k)a} \frac{[e^{ikx} + Q e^{i4K-kx}]}{1 - Q^2 e^{i4Ka}} . \tag{45}$$

Expanding the denominators in (43) and (44) enables us to interpret the solution in terms of multiple reflections inside the slab.

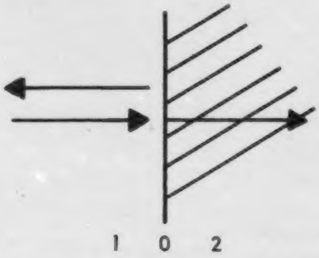


Figure 15-8c

If we are dealing with a single interface at $x = 0$ as in Figure 15-8c, then we obtain simply

$$E_1 = e^{ikx} - \frac{1 + Q}{2Q} e^{-ikx} , \tag{46}$$

$$E_2 = \frac{Q - 1}{2Q} e^{iKx} . \tag{47}$$

(See Exercises 15-8, No. 5.)

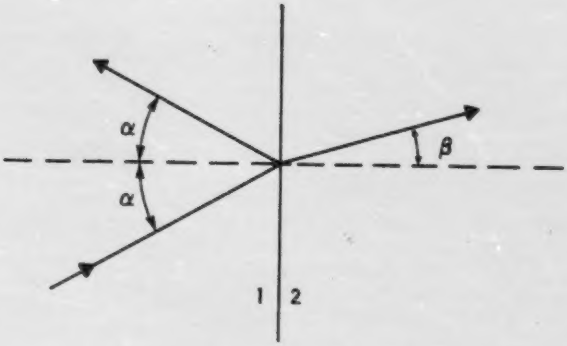


Figure 15-8d

The results and all the above may be generalized by inspection to an arbitrary angle of incidence α as in Figure 15-8d. Thus in Equations (32) and (33) we may replace k^2 by $k^2 \cos^2 \alpha$ and K^2 by $K^2 \cos^2 \beta$ and obtain the same final functions in terms of the new constants $k \cos \alpha$ and $K \cos \beta$, e.g., (46) becomes

$$e^{ikx \cos \alpha} - Q' e^{-ikx \cos \alpha} , \tag{48}$$

$$Q' = \frac{Z' - 1}{Z' + 1} , \quad Z' = \frac{KA \cos \beta}{k \cos \alpha} ,$$

and (47) becomes

$$(1 - Q')e^{iKx \cos \beta} . \tag{49}$$

We may now multiply (48) and (49) by the same factor $e^{iky \sin \alpha}$. This converts (48) to

$$E_1 = e^{ikx \cos \alpha + iky \sin \alpha} - Q'e^{-ikx \cos \alpha + iky \sin \alpha} = E_i(\alpha) - Q'E_i(\pi - \alpha) , \tag{50}$$

where $E_i(\alpha)$ is a plane wave incident at an angle α , and $E_i(\pi - \alpha)$ is its mirror image in the plane $x = 0$. From (49), multiplication by $e^{iky \sin \alpha}$ gives

$$(1 - Q')e^{iKx \cos \beta + iky \sin \alpha} . \tag{51}$$

If we require that

$$k \sin \alpha = K \sin \beta \quad , \quad \text{i.e.,} \quad \sin \alpha = \frac{K}{k} \sin \beta = \mu \sin \beta , \tag{52}$$

then (51) equals

$$E_2 = (1 - Q')e^{iKx \cos \beta + iKy \sin \beta} \tag{53}$$

which is a plane wave traveling at an angle β . Equation (52), which we recognize as "Snell's Law" [S] , is thus an artifice for converting the solution of a one-dimensional problem to the corresponding two-dimensional solution; it insures that corresponding wave fronts match at the surface.

Integral Representations. In Sections 15-6 and 15-7 we used Huyghens' principle in order to represent the total scattered field as the integral of "wavelets" arising from a distribution of elementary sources. To round out the previous intuitive discussion we should indicate how such forms follow from the present mathematical model [I] , [II] , and [III] . If we had

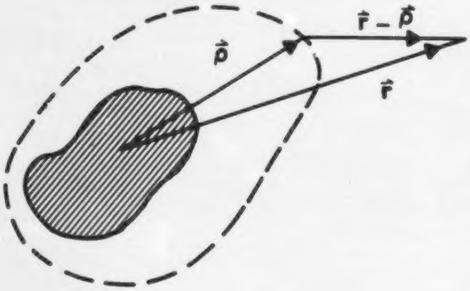


Figure 15-8e

available a theorem of Gauss (which relates certain surface and volume integral forms), we could prove that scattering functions E_s satisfying [I] and [III'] can be represented in terms of elementary sources H as

$$E_{sm}(\vec{r}) = \int_S [H_m(k|\vec{r} - \vec{\rho}|)\frac{\partial E(\vec{\rho})}{\partial n} - E(\vec{\rho})\frac{\partial}{\partial n} H_m(k|\vec{r} - \vec{\rho}|)]dS(\vec{\rho}) \ , \tag{54}$$

where $\vec{\rho}$ is a point on a surface $S(\vec{\rho})$ as in Figure 15-8e that incloses the scatterer but excludes the observation point $\vec{r}$, and where $\frac{\partial}{\partial n}$ is the outward normal derivative. The function $E(\vec{\rho}) = E_i(\vec{\rho}) + E_s(\vec{\rho})$, the total field at $\vec{\rho}$, and its normal derivative, are weighting factors for the surface distribution of elementary sources $H_m(|\vec{r} - \vec{\rho}|)$ and $\frac{\partial}{\partial n} H_m$ which radiate from $\vec{\rho}$ to $\vec{r}$.

The elementary sources are essentially those we worked with in earlier sections. Thus in three dimensions,

$$H_3(kR) = -\frac{e^{ikR}}{4\pi R} \ , \quad R = |\vec{r} - \vec{\rho}| = \sqrt{(x - \xi)^2 + (y - \eta)^2 + (z + \zeta)^2} \tag{55}$$

and in one dimension

$$H_1(kR) = \frac{e^{ikR}}{i2k} \ , \quad R = |x - \xi| \ . \tag{56}$$

In two dimensions, for argument $kR \gg 1$, we have

$$H_2(kR) \sim \frac{e^{-i\pi/4}}{4i}\sqrt{\frac{2}{\pi kR}}\, e^{ikR} \ , \quad R = \sqrt{(x - \xi)^2 + (y - \eta)^2} \tag{57}$$

which is the form we worked with previously; the fact that (57) holds only for large values of the argument accounts for the restriction $kR \gg 1$ that we mentioned for the strip and cylinder problems. For small values of kR , the elementary source in two dimensions behaves quite differently:

$$H_2(kR) \sim \frac{1}{2\pi} \ln kR \ , \quad kR \sim 0 \ . \tag{58}$$

Its exact representation is given by

$$H_2(kR) = \frac{1}{4i} H_0^{(1)}(kR) \ , \tag{59}$$

where $H_0^{(1)}$ (which is known as Hankel's function of the first kind of order zero) is the special solution of (20) for two dimensions corresponding for angle independent outgoing waves, i.e., it plays the same role in two-dimensional propagation as $\frac{e^{ikR}}{R}$ and $e^{ik|x|}$ play in the other cases. If we specialize $S(\vec{\rho})$ to the surface of the scatterer itself, then we can use such surface conditions as [II] to obtain equations (integral equations) for the unknown values of $E_s(\vec{\rho})$ and $\frac{\partial E_s(\vec{\rho})}{\partial n}$; for simple surfaces, the procedure is analogous to that we followed for the slab.

Although we will not prove (54), we will show how to obtain the approximate forms we worked with in earlier sections. Thus if we specialize $S(\vec{\rho})$ to the scatterer's surface and use the boundary condition [IIb] that $\frac{E(\vec{\rho})}{\partial n} = 0$, we reduce (54) to

$$E_s(\vec{r}) = -\int E(\vec{\rho})\frac{\partial}{\partial n} H_m(k|r - \rho|)dS(\vec{\rho}) \; . \tag{60}$$

In particular, in two dimensions and $k|\vec{r} - \vec{\rho}| \gg 1$, we use (57) in (60) to obtain

$$E_s(\vec{r}) \sim -\frac{e^{-i\pi/4}k}{4}\sqrt{\frac{2}{\pi k}}\int E(\vec{\rho})\,\frac{e^{ik|\vec{r}-\vec{\rho}|}}{\sqrt{|\vec{r}-\vec{\rho}|}}\,\frac{\hat{\rho}\cdot\hat{n}}{\rho}\,dS(\vec{\rho}) \tag{61}$$

which is of the required form [IIIc] . If we knew the field $E(\vec{\rho})$ on the scatterer's surface, we could obtain the scattered field $E_s(\vec{r})$ by integration. If we do not know the field (and it is only for very simple shapes that $E(\vec{\rho})$ is known exactly), then we may seek heuristic physically motivated approximations.

In particular if the scatterer is very big compared to wavelength, then it is plausible to approximate $E(\vec{\rho})$ by elementary geometrical optics considerations. Following essentially Kirchhoff, one approximates the total field $E(\vec{\rho})$ on the "lit side" of the scatterer by twice the incident value E_i , and by zero on the "dark side." Thus if we substitute

$$E(\vec{\rho}) \approx E_i(\vec{\rho}) \text{ on lit side; } \; E(\vec{\rho}) \approx 0 \text{ on dark side} \tag{62}$$

into (61) we obtain the general case of the integrals we considered in Sections 15-7 and 15-8, i.e., for the strip with $dS(\vec{\rho}) = d\eta$ and the circular cylinder with $dS(\vec{\rho}) = a d\phi$.

From (54) we could also construct the volume integral of elementary scatterers that we worked with previously for the case of a partially transparent sphere. First we specialize (54) to $\vec{\rho}$ on the scatterer and then use the transition conditions [IIc] to replace the external surface fields $E(\vec{\rho}) = E(k,\vec{\rho})$ and $\frac{\partial}{\partial n} E(\vec{\rho})$ by the corresponding internal fields $E(K,\vec{\rho})$

and $\frac{A\partial}{\partial n}E(K,\vec{\rho})$ where $K = k\mu$ is the internal wave number. We then use the same theorem of Gauss to convert the resulting surface integral to an integral over the volume of the scatterer. In particular for constant μ and $A = 1$, we would obtain

$$E_s = (k^2 - K^2)\int_V H(k|r - \rho|)E(K,\vec{\rho})dV(\vec{\rho}) \tag{63}$$

where V is the volume of the scatterer. If we add E_i to both sides of (63) we obtain an integral equation for E which can be solved for simple shapes. We shall not prove (63), but we will show how this rigorous result leads to the previous approximation of Section 15-6, Equation (38).

For tenuous scatterers in the sense $K^2 = k^2\mu^2 \sim k^2$ Rayleigh replaced the unknown internal field $E(\rho)$ in (63) by the incident wave:

$$E(K,\vec{\rho}) \approx E_i(k,\vec{\rho}) = e^{ik\zeta} . \tag{64}$$

If we substitute (64) into (63) and specialize to three dimensions by using (55), we obtain

$$E_s \approx \frac{k^2(\mu^2 - 1)}{4\pi}\int \frac{e^{ik|\vec{r}-\vec{\rho}|}}{|\vec{r} - \vec{\rho}|}\, e^{ik\zeta}\, dV(\vec{\rho}) , \tag{65}$$

which is the more complete version of the form with which we worked previously in Section 15-6, Equations (38)ff, Equation (65) gives directly the result that the flux $(|E_s|^2)$ is inversely proportional to λ^4 for scatterers whose length dimensions are small compared to λ .

It should be stressed that such approximations as (62) and (64) are adequate only for limited ranges of the parameters. However, within their limintations, they provide useful and instructive explicit results for problems that cannot be solved rigorously.

In concluding this chapter, we should also reiterate and stress that we have covered merely a selected sequence of topics in wave physics. The wave equations that we introduced in ad hoc fashion are generated more systematically in physics courses by combining first order differential equations (Maxwell's equations for electromagnetics), which arise partly from interpretation of the physical experiments of Faraday and others. Although we made "light" the theme for much of the development, we have not covered an essential aspect that distinguishes wave models for light from the models used for sound: light, and all electromagnetic waves, must also be characterized by polarization; this requires in general that we deal with vector wave functions with amplitudes perpendicular to the direction of propagation instead of the scalar functions we have considered. Our discussion of light was in no sense meant to be comprehensive and there are many phenomena involving light that are not described by a wave model at all. In illustrating different applications of calculus, we have used light as a vehicle for an introduction to wave physics, not only because we have many visual experiences to draw on, but because the adequacy of a wave model for such phenomena was far from obvious to the early investigators (and not particularly obvious even to us without some careful observations). For water waves, the appropriateness of the mathematical model would have been clear from the start, and even for sound waves the intuition leads relatively directly from the visible waves on stringed instruments and on drumheads to waves in air. Thus in discussing light, we could introduce key topics leading to the development of the wave model essentially in their historical order, and thereby indicate the greater generality and the greater economy of the wave model over the earlier collection of special "laws of nature." However, the initial reservations that "light" is neither wave nor particle, and that only certain classes of phenomena involving light are adequately described by a wave model should not be lost sight of. Light is one of the most complex characters in the ABC of mathematical physics.

Exercises 15-8

1. Verify that the plane waves

$$U = e^{\pm ikx\cos\alpha + iky\sin\alpha - i\omega t}$$

are solutions of (9) as claimed.

2. Show that any sufficiently differentiable function of the form

$$U = F(vt - x) + G(vt + x)$$

satisfies Equation (8).

3. Verify that a solution of the wave equation (10) has the form (17), namely

$$E(x,y,z,t) = f(x,y,z)g(t) ,$$

only if

$$\frac{1}{f(x,y,z)}\,\Delta f(x,y,z) = \frac{1}{v^2 g(t)}\,\frac{d^2 g(t)}{dt^2} = \text{const.}$$

4. Obtain (43) and (44) as solutions of the linear system (38) - (41).

5. Derive Equations (47) and (48).

INDEX

absolute convergence of integrals, 664
absolute value, 254, 274
acceleration, 2, 409
Al, the story about, 497, 659
angle
  of inclination, 30
  of intersection of curves, 491
angular frequencey, 962
angular momentum, 44
antiderivative, 427
Arago bright spot, 923, 986
arccos, 144
arclength, 43, 384, 385, 650ff
arcsin, 143
arctan, 144
area, 11, 367ff
  between the graphs of two functions, 397
  intuitive concept, 11
  invariance of, 56
  of a standard region, 13f
area function
  additive property, 367
  order property, 367
asymptote, 230f
  horizontal, 232, 234
  oblique (slant), 233, 234
  vertical, 232, 234
attenuation, 502

base of a power, 445
Binomial Theorem, 338, 857
binormal, 76
bound
  lower, 261
  upper, 261
boundary conditions, 1006
bounded growth, 512
bounded set of points, 261
  greatest lower bound, 266
  least upper bound, 265
bounded variation, 649
braking coefficient, 513
Buniakowsky's-Schwarz inequality, 403

cardioid, 48
catenary, 489
Cauchy Condensation Test, 872
Cauchy Convergence Theorem (Th. 14-2k), 859
Cauchy Criterion for convergence, 864
Cauchy form of remainder (Taylor's Theorem), 825
Cauchy product, 880
Cauchy sequence, 859
Cauchy's inequality, 253, 403
caustic
  line, 910
  point, 910
  surface, 910
  virtual, 916
caustic epicycloidal, 917
caustics
  field on, 999
  rainbow, 930
center of curvature, 62
center of mass, 51
centoid, 14
central force, 44
Ceva's Theorem, 20
Chain Rule, 149
change of origin, 18
characteristic equation, 605
chronaxie, τ , 509
circular frequency, 14
circular functions, 137, 303
  analytical definition of, 653
  derivative of, 137
  derivative of inverse, 143
  differential equations for, 471
Clairaut Equation, 65
cluster points, 861
collinearity of vectors, 9
competition, 512
completeness of the real number system, 263
component, 5
  parallel, 33
  perpendicular, 33
  of a vector, 5
composite function
  derivative of, 149
composition
  of functions, 102, 285
  of translations, 6
conditinuous dependence on initial data, 593
conic section, 313
  directrix, 313
  eccentricity, 313
  focus, 313
conservation of angular momentum, 44
  partial, 52
conservation of energy, 11, 31
constant function, 274
constant phase, 963
constrained extreme value problems, 213

- constraint, 213
- continuity, 91ff
  - at a point (definition), 93
  - of composition function (Th. 3-6e), 102
  - of a differentiable function (Th. 3-6d), 101
  - on the interval, 108
  - intuitive idea, 162
  - of inverse function (Th. 3-6f), 104
  - piecewise, 589
  - of product of continuous functions (Th. 3-6b), 99
  - of quotient of continuous functions (Th. 3-6c), 100
  - of rational combination, 99
  - of a strongly monotone function (No. 16), 115
  - of sum of continuous function (Th. 3-6a), 99
- continuous
  - dependence on initial data, 602
  - piecewise, 589
- continuous function
  - boundedness, 345
  - integrability, 648
  - monotone between extrema, 178
- convergence
  - absolute, 664, 873
  - conditional, 873
  - pointwise (Def. 14-6a), 883
  - radius of, 891
  - of sequences, 852
  - uniform (Def. 14-6b), 885
- convex function, 206
  - boundedness of, 212
  - continuity of, 212
  - differentiability of, 212
- convex point, 945
- convex set, 207
- convexity, 206
  - flexed downward, 207, 234
  - flexed upward, 207, 208, 234
- coordinate representation, 5
- coordinates
  - polar, 308
  - properties independent of (see invariance), 2
- coplanar vectors, 7
- Cornu spiral, 47
- cosine, (see circular function), 137
  - integral, 635
- cover
  - of a closed interval, 645
  - of an interval, 645
- critical damping, 23
- cross product
  - right hand rule for, 27
  - of two vectors, 26
- curvature, 58
  - as characterization of a plane curve, 67
  - center of, 62
  - invariance of, 58
  - radius of, 59
  - sign of, 59
- curve
  - arclength of, 43
  - Frenet-Serret equation for, 77
  - motion along, 35
  - parametric representation of, 40
  - piecewise smooth, 40
  - principle normal, 48
  - rectifiable, 651
  - simple closed, 51
  - vector tangent, 41
- cycle, 14
- cycloid, 47

- damped oscillation, 14
- decay, 499
- decay coefficient, 499
- decimal
  - periodic, 267
- decomposition into partial fraction, 563
- decreasing, (see increasing), 299
- decreasing function
  - (see increasing function) 110,234, 299
  - Weakly, 196, 299
- definite integral, 427
- derivative, 5, 27ff
  - of a^x, 466
  - of arccos x, 147
  - of arcsin x, 147
  - of arctan x, 147
  - of a circular function, 137ff
  - of a composite function, 149
  - of compositions (Chain Rule) (Th. 4-6), 149
  - of a constant, 118
  - of cos x, 139
  - of cot x, 139
  - D_x, 117
  - of e^x, 465
  - of an exponential function, 465
  - at an extremum, 173
  - of f', 117
  - of $f : x \to c$, 118
  - of $f : x \to x$, 118
  - of $f : x \to x^c$, 118
  - of $f : x \to \sqrt{x}$, 118

  of $f : x \rightarrow \frac{1}{x}$, 118
  of $f : x \rightarrow |x|$, 118
  of a fractional power, 134
  of a function at a point, 49
  of an implicitly defined function, 161
  of an integral power, 125
  intermediate value property of, 195
  intuitive concept, 5ff
  of an inverse, 131
  of an inverse circular function, 143ff
  of inverse of differentiable function (Th. 4-3), 132
  left-sided (No. 7), 121
  Leibnizian notation, 156
  of linear combination (Th. 4-2a), 120
  of log x, 465
  of the logarithmic function, 465
  of a monotone function, 196
  notations for, 154ff
  of a polynomial (Th. 4-2c, Cor. 2), 125
  of polynomial of differentiable function (Th. 4-2c, Cor. 3), 125
  of a power function, 465
  power rule for positive integers (Th. 4-2c), 125
  of a product (Th. 4-2b), 122
  of a quotient of differentiable function (Th. 4-2d, Cor. 1), 127, 128
  of a rational function (Th. 4-2d, Cor. 2), 129
  of reciprocal of differentiable function (Th. 4-2d), 128
  of right-hand and left-hand, 121
  sign test for, 198
  of sin x, 139
  successive higher, 159
  of tan x, 139
  of a vector function, 38
differential equation, 429
  e^x, (Th. 3-5a), 471
  linear first order, 550ff
  linear second order, 603ff
  separable, 621ff
  sin x, cos x, (Th. 8-5b), 472
differential operator, 590
differentiation, 117ff
  linearity of, 120
  partial, 1002
direct sum of linear vector spaces, 12
direction angle, 30
Dirichlet integral, 664
discontinuity
  point of, 94
displacement, total, 408
domain of a function, 269
dot product of two vectors, 22
dynamics, 3

e, 461, 480
  irrationality of, 479
  properties of, 477
edge diffracted rays, 920
eikonals, 917
ell, 881
ellipse, 313
  focal chord, 314
  latus rectum, 314
energy, 11
  density, 503
envelope, 63
epicycloid, 75
epsilonics, 67f
equilibrium, 33, 39
  stability of, 40
equivalent parametrization, 44, 49
escape velocity, 49
estimate
  lower, 255
  upper, 255
Euclid's Principle
  of propagation, 904
  of reflection, 904
Euler's constant, 861
Euler's Method , 842
Euler polygon, 843
evolute, 62
  as envelope of normals, 63
exponent, 445
  definition of zero exponent, 446
  general laws for negative integers, 446
  general laws for positive integers, 445
  rational exponents, 447
exponential function, 447, 458ff
  derivative of, 448
  differential equation for, 471
  inverse function, 448
  relative magnitude, 463
exponentially damped sinusoid, 607
Extreme Value Theorem (Th. 3-7b), 109, 345ff
  proof, 347
extremum, 173
  derivative at, 173
  isolated, 199

  is dated, 199
  local, 176, 181, 200
  location of, 73ff
  on open interval (Lemma 5-2), 178
  relative, 176
  second derivative test for, 205
  sign of derivative test, 198
  strong, 419

factorial
  definition of, 328
  estimates of, 481
Faraday, 1015
Fermat's Principle of Refraction, 928
field, 245
  ordered, 249
First Comparison Test for Convergence (Th. 14-3c), 865
flexure, 206ff
flux density, 941
flux principle, 939
folium of Descartes, 228
force, 3
forcing term, 591
Fourier Series of f, 890
fractional part, 278
Fraunhofer diffraction, 968
free vibration problems, 1006
Frenet-Serret equations, 77
frequency, 14
  driving, 16
  natural, 16
Fresnel diffraction, 978, 981
Fresnel integral, 981
friction
  linear, 14
  sliding, 27
function, 269ff
  absolute value, 95, 274
  of bounded variation, 649
  circular, 303
  composite, 285ff, 286
  continuous and nowhere differentiable, 111, 352
  convex, 206
  even, 276
  exponential, 447, 458
  hyperbolic, 485
  implicitly defined, 161, 359ff
  increasing, 299
  increasing on the right, 349
  integer part, 57, 275
  inverse of, 290, 300
  limit of, 58ff
  linear combination of, 78
  logarithmic, 448
  monotone, 298
  odd, 276
  one-to-one, 290
  periodic, 277
  power, 459
  rational combination of, 78
  signum (sgn), 61, 62, 276
  strongly monotone, 299
  weakly increasing, 299
function definition, 269
  circular, 137, 303
  constant, 274
  explicitly defined, 162
  identity, 274
  implicitly defined, 161
  inverse circular, 143f
fundamental, 98
fundamental set of vectors, 28
fundamental solution, 594
Fundamental Theorem of Calculus, 425

Galilean Principle of Relativity, 8
global property of f, 169
gnomon, 881
graph
  of a function, 272
  sketching, 229, 233
Greek Alphabet, 244
Green's function, 616
Grimaldi, 922
growth coefficient, 497, 513

half-life, 499
harmonic oscillator, 14
heaviside pulse, 951
Heine-Borel Principle, 645
helix, 48
Helmholtz's equation, 1004
Hero's Principle (of reflection), 906, 909
homogeneous equation, 591
Hooke's law, 13
Huyhen's Principle, 951
hyperbola, 313
hyperbolic functions, 485
  cosh x, 485
  derivative of, 485
  differential equation for, 489
  geometrical interpretation, 487
  inverse of, 488
  sinh x, 485
  tanh x, 485
hyperbolic sector, 487
hypocycloid, 75
  of four cusps, 44

- Identity Function, 274
- inclined plane, motion on, 26
- image, 270
- implicit differentiation, 162
- Implicit Function Theorem, 361
- improper integral, 578
  - convergence of, 661
- increasing function, 234, 299
  - continuity of, 110
  - weakly, 196, 299
- indefinite integral, 427
- inequality
  - strong, 249
  - weak, 249
- inertial coordinate system, 4
- inflection point, 230
- initial value, 497
- initial value problem, 430, 592
- integer part, 275
- integrability
  - of a continuous function, 645ff, 648
  - of a linear combination of monotone functions, 425ff
  - of a monotone function, 378
  - of a piecewise monotone function, 415
- integral, 11, 377
  - absolute convergence of, 664
  - of a continuous function, 648
  - convergent, 582
  - definition, 377
  - estimate of, 437
  - existence of, 638ff
  - Existence Theorem (Th. 6-3a), 378
  - geometric properties, 388
  - improper, 578
  - intuitive concept, 11ff
  - limit of Tiemann Sum, 383, 643
  - of a linear combination, 393ff
  - of monotone function (Th. 6-3b), 379
  - of a motion, 11
  - nonnegative function, 401
  - operator, 617
  - of a periodic function, 401, 571
  - of a rational combination of circular functions, 550
  - of a symmetric function, 570
  - test for convergence (Th. 14-3d), 866
- integrals
  - convergent, 582
  - definite, 570, 427
  - definition, 581
  - divergent, 582
  - improper, 578
  - symmetric, 571
- integration, 535
  - of constant times integrable function, 394
  - formal, 433
  - of linear combination of integrable functions, 393
  - linearity of, 393, 430
  - by parts, 554
  - of a polynomial, 633
  - of rational functions, 563
  - special reductions, 573
  - substitution of circular functions, 546
  - Substitution Rule (Th. 10-2), 540
  - of sum of integrable functions, 395
  - by summation, 633
- interference, 966
  - constructive, 968
  - destructive, 968
- interior point of an interval, 259
- Intermediate Value Theorem (Th. 3-7a), 109
  - proof, 350f
- interval, 259
  - closed, 259, 109
  - interior point of, 259
  - inverse function, 131, 291f
  - length of, 259
  - midpoint of, 259
  - open, 109, 259
- interpolation, linear, 191
- invariance, 2
  - of area, 56
  - of curvature, 58
- inverse
  - of a function, 290
- Inverse Circular Functions, 143
  - derivative of, 147
- inverse function
  - derivatives of, 131
- inverse hyperbolic functions, 488
  - derivative of, 489
- inverse square force, 47
- involute, 62
  - geometrical construction of, 73
- iteration schemes
  - alternating, 817
  - approximate x for f(x) = 0, 809
  - convergent, 809, 814
  - square root, 808

- Jordan curve, 51

- Keller, J.B., 920
- Kelvin, 979
- Kepler-Lambert Principle, 939f
- Kepler's First Law, 47

- Kepler's Second Law, 45
- Kepler's Third Law, 50
- kinematics, 2
- kinetic energy, 11, 31
- Kirchhoff, 1014

- Lagrange
  - form of remainder (Taylor's Th.), 825
  - rule of variation of parameters, 615
- latent period, 509
- law of inertia, 3
- Law of the Mean, 186ff, 190
  - generalized, 76
- Law of Mass Action, 516
- Least Upper Bound Principle, 265
- Leibniz, 156
- Leibnizian notation
  - for derivatives, 156
  - for integral, 383
- Leibniz's Test for Alternating Series (Th. 14-4b), 874
- lemniscate of Bernouli, 313, 359
- length, 4
  - of an interval, 259
- limit
  - of a constant, 79
  - of f at a, 58f
  - of a function, 55ff
  - of a function (definition), 59
  - of a function at infinity, 231
- limit inferior, 862
  - intuitive concept, 6
  - of a linear combination, 81
  - of a product, 82
  - of a quotient, 85
  - right- and left-sided, 90, 578
  - $\frac{\sin x}{x}$, 138
  - of a sum, 80
  - superior, 861
- limits, 55
- limit theorems
  - constant function
    - (Th. 3-4a), 79
    - (Th. 14-2a), 852
  - constant multiple of a function
    - (Th. 3-4b), 79
    - (Th. 14-2b), 852
  - linear combination of functions
    - (Th. 3-4c, Cor.), 81
    - (Th. 14-2c, Cor.), 852
  - nonnegative function
    - (Lem. 3-4, Cor. 2), 84
    - (Lem. 14-2, Cor. 2), 853
  - product of functions
    - (Th. 3-4d), 82
    - (Th. 14-2d), 853
  - rational function
    - (Th. 3-4e, Cor. 2), 86
    - (Th. 14-2e, Cor.), 853
  - reciprocal of a function
    - (Th. 3-4e), 85
    - (Th. 14-2e), 853
  - Sandwich Theorem
    - (Th. 3-4f, Cor. 1), 86
    - (Th. 14-2f, Cor. 1), 853
  - Squeeze Theorem
    - (Th. 3-4f, Cor. 2), 87
    - (Th. 14-2f, Cor. 2), 854
  - sum of functions
    - (Th. 3-4c), 80
    - (Th. 14-2c), 852
- line
  - vector equation of, 13
- linear approximation, 224f
- linear approximation of f, 223
- linear combination, 78
- linear differential equation of first order, 590
  - forcing term, 591
  - fundamental solution, 594
  - general solution, 594
  - initial value problem, 592
  - nonhomogeneous equation, 595
  - reduced equation, 591
- linear differential equation of second order, 603
  - homogeneous equation, 604
  - superposition principle, 604
- linear friction, 14
- linear interpolation, 191
- linear operator, 591, 601
- linear vector, space, 9
- local property, 169
  - of a function, 108
- logarithm, 448
  - base e, 461
  - base 10 (common), 461
- logarithm
  - derivative, 449
  - estimates for, 456
  - function, 448
  - as an integral, 452
  - of a product, 453
  - of a quotient, 453
  - with any base, 461
- logistics equation, 513
- Lorentz force, 18
- lower sum over σ , 376

mapping, 270
mass, 5
mathematical induction, 369ff, 319
    first principle, 323
    second principle, 327
maximum (see extremum), 173
    local, 177, 181, 198, 205, 234
    of a set, 255
maxwell, 1015
mean life-time, 499
Mean Value Theorem of integral calculus, 402
Menelaus's Theorem, 21
method of equated coefficients, 566
minimum (see extremum), 173
    local, 177, 181, 198, 205, 234
    of a set, 255
model
    for decay, 499
    for growth, 497
momentum, 6
monochromatic light, 959
Monotone Convergence Theorem (Th. 14-2h), 856
monotone function, 298ff, 196, 415
    derivative of, 196, 204
    integrability of, 378
    inverse of strongly monotone function (Th. A2-4), 300
    linear combinations of, 415
    piecewise, 415
    sectionally, 415
    strongly, 178, 299

neighborhood, 58, 260, 259
    deleted, 58, 260
    of infinity, 587
    radius of, 260
Nested Interval Principle, 265
Newton, 156
Newton's First Law, 4
Newton's Second Law, 5
Newton's Third Law, 6
Newton's Method, 813
nonhomogeneous equation, 595
norm of the partition, 379
normal
    line, 226
    to a plane, 31
    at a point, 226
    principle to a curve, 74
notation
    D_x, 117
    Δ (difference), 156
    Δ (increment), 149
    $\frac{dy}{dx}$, 156
    f', 117
    Leibnizian, 156
null vector, 8
number
    irrational, 263
numberical integration, 829f
    rectangle rule, 829
    trapezoid rule, 830
    Simpson's Rule, 832
    Stirling's Formula, 835

operator
    differential, 590
    linear, 591, 601
orthogonal trajectories, 622, 917
osculating circle, 62

parabola, 313
parallelism
    of directed segments, 4
parallelogram
    law for vector addition, 7
parameter, 44
parametric representation of a curve, 40
paraxial rays, 915
particle, 2
    charged, 18
particular solution, 595
partition, 376
partition of [a,b], 376
pendulum, 32
    cycloidal, 36
    period of, 34
    spherical, 52
period
    of a decimal, 267
    of a function, 277
    fundamental, 98, 278
    of planetary motion, 50
periodic waves, 959
phase, 599, 960
    method of stationary, 979, 989
phase lag, 16
pi
    definition of, 655
Picard's Method, 814
piecewise continuous, 589
piecewise monotone functions, 415
    integrability, 415
planar source, 942
plane
    characterization by dot product, 25
    of incidence, 905
    normal to, 31
    vector equation of, 16
planetary motion, 47
point of inflection, 230, 234

Poisson bright spot, 923
polar axis, 308
polar coordinates, 308
polarization, 1015
polynomial
  derivative of, 125
  number of zeros, 188
position vector, 8
potential energy, 11, 31
potential function, 38
potential well, 38
power, 445, 459
power series, 891
primitive of f, 427
principal normal, 74
Principles of Archimedes, 266
product, of functions, 285
p-test for convergence
    (Th. 14-3e), 867

radiation
  primary, 948
  secondary, 948
radioactive decay, 500
radius of curvature, 59
radius of a neighborhood, 260
radius vector, 308
range of a function, 269
ratio test for convergence
    (Th. 14-3g), 868
rational combination, 78, 99
Rayleigh-Born scattering, 970f
reaction rate, 516
real numbers, 245
  algebraic properties of, 245
  order relations, 249
real number system, 245ff
  completeness, 263ff, 645
rectifiable, 651
recurrence relations, 558
recursion, 238
  definition by, 328
  reduced equation, 591
  reflected ray system, 905
reflectors
  parabola, 913
  semicircle, 914
refraction
  index of, 927
resistance
  of air, 11, 23
resonance, 16
restriction, 297
  of a function, 297
rheobase, 508
Riemann sum, 381
  limit of, 383
  upper, 384
rocket
  motion of, 21
Rolle's Theorem
    (Lemma 5-3), 187
root
  principal, 301
  test for convergence
    (Th. 14-3h), 869

Sandwich Theorem, 86
scalar, 9
scatterer
  point, 1008
  slab, 1009
scattering, 503
  amplitude, 970
  coefficient, 503
Second Comparison Test for convergence
    (Th. 14-3f), 868
second derivative, 205
semicubical parabola, 42
separable differential equation, 621
Separation Axiom, 263
separation constant, 1004
sequence, 851
sequence of partial sums, 863
series, 863
set
  convex, 207
signum, 276
simple closed curve, 52
  area enclosed by, 52
  orientation of, 52
Simpson's Rule, 832
sine (see circular function), 137
slope, 5, 27ff, 30
smooth, 54
  piecewise, 54
Snell's Law of Refraction, 927
solid of revolution, 405
spherical pendulum, 52
Squeeze Theorem, 87
stability, 40
standard region, 13
  lower bound, 370
  upper bound, 370
steady state, 599
Stirling's Formula, 482, 835
stratified medium, 933
strongly monotone, 299
Substitution Rule, 428
  for integrals, 540
substitutions
  of circular functions, 546
  of hyperbolic functions, 553
sum
  of functions, 285
  Riemann, 381

    upper, 377
    of vectors, 7
Sum Notation, 333, 371
summation, 339
Superposition Principle, 604
supremum, 265
symbol
    max $\{r_1,r_2,\ldots,r_n\}$, 255
    min $\{r_1,r_2,\ldots,r_n\}$, 72, 74, 76
        83, 255
symmetry, 570

tangent, 29
    to the curve, 223
Taylor expansion, 823
Taylor series, 823
Taylor's Theorem
        (Th. 13-3), 820
tolerance, 6
tolerance ϵ (error), 32, 63
transient state, 599
transition conditions, 1006
translation, 4
translations
    composition of, 6
triangle inequality, 255
triple scalar product, 31
trivial solution, 591
trochoid, 21
truncation error, 845

unit line source, 940
unit point source, 940
upper sum over σ, 377

variable, 274
variation of parameters, 596
vector
    component of, 5
    coordinate representation of, 5
    coplanar, 17
    cross product of, 26
    definition of, 4
    dot product of, 22
    length of, 4
    linear dependence of, 19
    multiplication by a scalar, 9
    position, 8
    triple scalar product of, 31
velocity, 2
    average, 30, 42
    instantaneous, 42
volume of solid of revolution, 405

Wallis's Product for $\frac{\pi}{2}$, 575, 836
wave equation, 1002
wave length, λ, 925
wavelet, 951
wave surface, 951
weakly, increasing, 299
Weierstrass, 111
Weierstrass Function, 352
Weierstrass M-Test
        (Th. 14-6c), 887
well-posed problem, 593
work, 31

Young's Principle, 960

zero of order k, 826